KU LEUVEN

ARENBERG DOCTORAL SCHOOL
Faculty of Engineering Science

Automated sleep analysis: from the hospital to the home environment

Elisabeth Heremans

Supervisors:
Prof. dr. ir. M. De Vos
Prof. dr. W. Van Paesschen

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Engineering Science (PhD): Electrical Engineering

October 2024

Automated sleep analysis: from the hospital to the home environment

Elisabeth HEREMANS

Examination committee:
em. Prof. dr. ir. O. Van der Biest, chair
Prof. dr. ir. M. De Vos, supervisor
Prof. dr. W. Van Paesschen, supervisor
Prof. dr. D. Testelmans
Prof. dr. M. Blaschko
Prof. dr. ir. C. Phillips
(ULiège)
Prof. dr. M. Van Gilst
(TU/e)

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Engineering Science (PhD): Electrical Engineering

October 2024

© 2024 KU Leuven – Faculty of Engineering Science
Uitgegeven in eigen beheer, Elisabeth Heremans, Kasteelpark Arenberg 10 box 2446, B-3001 Leuven (Belgium)

Alle rechten voorbehouden. Niets uit deze uitgave mag worden vermenigvuldigd en/of openbaar gemaakt worden door middel van druk, fotokopie, microfilm, elektronisch of op welke andere wijze ook zonder voorafgaande schriftelijke toestemming van de uitgever.

All rights reserved. No part of the publication may be reproduced in any form by print, photoprint, microfilm, electronic or any other means without written permission from the publisher.

Preface

The first words of this thesis were written on a train to Ghent, where I had to go plead with the NMBS to get my phone back after accidentally leaving it on the train one day before. This happened on my third commute from Brussels to Leuven. Fortunately, the phone rescue operation was successful. Unfortunately, my beloved headphones weren't so lucky as to survive my four-year PhD journey, although they might be living their best life on Rhodes Island right now. You might start to wonder: 'how does someone so chaotic and forgetful manage to write a thesis?' I do not have the answer to that question, but I do know it definitely wouldn't have been possible without the unwavering support of many incredible people.

First, I would like to express my heartfelt gratitude to my promoter, Prof. De Vos, for his invaluable guidance and support over the past four years. This PhD journey has been an incredible learning experience, and I feel fortunate to have had you as my mentor. Even with your often packed schedule filled with meetings from morning to evening, you still consistently found the time to give me feedback whenever I needed it. I am also truly grateful for the freedom you gave me to travel and pursue projects and research stays beyond KU Leuven. Lastly, it was a real joy to finally play the violin together in the last month of my PhD. Maarten, thank you so much for your mentorship!

I would also like to extend my sincere gratitude to prof. Van Paesschen for the invaluable clinical insights you shared during our hospital meetings. Your guidance and perspectives were instrumental in helping me maintain a clinical viewpoint in my work. I am also grateful to prof. Buyse and prof. Testelmans for their support and collaboration on the elderly OSA sleep study, which provided an important dataset for validating my methods. Additionally, thank you, prof. Testelmans and prof. Blaschko, for engaging with my work as assessors throughout the years. Finally, I want to express my appreciation to my other jury members, prof. Phillips and prof. Van Gilst, for their critical questions that helped refine my thesis, and to prof. Van der Biest for chairing

my examination committee.

Additionally, I had the pleasure to collaborate with researchers at UZ Leuven and elsewhere, which enriched my PhD experience and made it even more interesting. Thank you for the nice collaborations, Nabeel Seedat, Astrid Devulder, Pascal Borzée, Letizia Micca, Ren'ee Proost, Jaiver Macea Ortiz, Laura Van den Bulcke, prof. Maarten Van Den Bossche, prof. Gilat, dr. Huy Phan, and prof. Mihaela van der Schaar.

These four years were made a thousand times better by being in a really wholesome and supportive group of fun-loving and cookie-eating people. Biomed, I honestly can't imagine a nicer group of colleagues to go through these four PhD-years than you guys! We bonded over so many coffee breaks, tons of food, sporty activities, and drinks in Leuven. First of all, thank you Konstantinos, Joran, and Lore, for sharing an office with me. Konstantinos, your overall chillness and genuine kindness make everyone feel at ease around you. Thank you for all of the fun chats, BBQs, and so much more. Joran, you were always ready to bounce off ideas and engage in thoughtful discussions. I loved our 33k bike and run (destroying the other team hehe). Lore, I'm really glad you got sorted into our office. As I pass on my desk to you, remember: with a great desk comes great responsibility. If you could take care of Amalia's one lonely plant that has managed to survive my PhD, I'd greatly appreciate it. I feel like it symbolizes the struggle for survival of a PhD.

To the Biomed crew that went to ICASSP 2023 in Rhodes, this was by far the most fun conference in my PhD and I was so happy to get to know you all better there. Cem, thanks for always bringing color into every room you step into, literally and figuratively, for organizing so many fun activities (dinners, museum trips, you name it), and for keeping the Biomed spirit alive in tough covid times. Miguel, I'm happy to almost be able to communicate with you in your native language now (BRAZIIIIL). I really enjoyed all of our *kopjes koffie* (I should say *xícaras de café*). Since you liked California so much, feel free to come by anytime! Nicolas, my most consistent running buddy! Thanks for always keeping the conversation going while I was out of breath ('no, no, the pace is really fine!' *dying inside*). Thomas, your cynical sense of humor never fails to spice up lunch time, and your competitive spirit always brought out the best in us at every sporty Biomed event. Christos, you're always willing to lend a hand and share your wise advice, which I imagine you've gathered over your millions of years working here. Nick, thanks for organizing the most amazing EEG workshop in the history of humankind with me.

Sorry, I'm still not finished with the Biomed list of people. Angeliki, you're one of the most lovable people I know, and I'm happy to call you my friend. Let's go to Meli sometime soon please (I need some bougatsa)! Simon, thank

you for sparking my enthusiasm for scicomm and always bringing such a lively atmosphere! Although I wish our common office wall had been a bit thicker. Iustina, you're such a friendly and genuine person, and I have no doubt you will be an incredible mom. I hope it's not too soon to give my heartfelt congratulations to both you and Teo! Charles, if there was a prize for most babies obtained during the PhD, you would win the gold medal. I wish you all the best with your family. Jingwei and Luis, I loved all the cool Chinese treats you brought on Fridays. Who knew I could like cookies stuffed with egg yolks? Nefeli and Lennert, I feel like I owe you a bar crawl or good party since I bailed last time. Let's make it happen soon! Marta, my fellow Protagonist, thanks for bringing so many beyblades and good vibes to the office. Guido, I never saw anyone explaining why dark mode is the only right VS Code color theme with so much Italian passion and hand gestures. I hope you finally convince Marta one day. Christof, I loved our fondue nights and I was so glad to find someone equally passionate about cheese as I am. Helene, thanks for being such a calm and lovely presence in the office, I was really glad to get to know you.

I also want to thank all the other Biomedians: Congting, Kenneth, Raphaël, Stefan, Bavo, and Michiel, for the good times, coffees, and conversations! And thank you to all past Biomedians who welcomed me when I started here in March 2020, right before tumultuous times: Mario, Amalia, Dries, Jonathan, Ofelie, and many more. Thank you, Maarten and Alexander, for leading this group and helping to create such a nice atmosphere here, and thank you, Sabine, for introducing me to Biomed when I was still a bachelor student. Thank you all for an unforgettable four years.

After long days spent contemplating neural networks and EEG signals, it's important to take time to relax and recharge. I want to thank my friends from school, orchestra, and university for helping with that vital part. Thank you for the chats, adventures, weekend trips, travels, shared hobbies, musical endeavors, and all the beautiful moments we have shared.

Dankjewel 'wij': Laura, Aster, Heleen, Ming-Sum, Ye, Nele, en Bona. Jullie zijn al zo'n 15 jaar de veilige vriendinnengroep waar ik altijd op kan rekenen. Aster, wij kennen elkaar al van de kleuterklas, en onze zelfde schommel heeft ons van toen af verbonden. Heleen, Ming-Sum, Nele, en Ye, vanaf Latijn-Grieks in klas 2A mijn onafscheidelijke vriendinnen. Van Hema-hangouts en supergenante fotoshoots op de speelplaats tot reizen samen en heel veel brunchen en nog zo veel meer. Ye, ook bedankt voor de muziek die ik met je heb mogen delen (Lamitié!). Laura, mijn achter-achter-achternicht, allerbeste klimmaatje en de persoon aan wie ik alles kan vertellen, dankjewel voor alles. En Bona, lieve lieve Bona, mijn gemis is enorm. Dankjewel voor je schaterlach die te pas en te onpas elke situatie opvrolijkte. Je blijft voor altijd een deel van mij en van 'wij'.

Dankjewel ook aan mijn vrienden van het USO, die het universiteitsleven zoveel mooier gemaakt hebben. Bedankt voor de muziek en de mooie vriendschap die eruit gevolgd is, Marjolijn, Isabelle, Manon, Katrijn, Katrien, Astrid, Sarah, Sebastiaan, Simon, Elias, Louis D., Robin, Jules, Lauranne, Ilka, Liesbeth en nog vele anderen. Ook dankjewel aan mijn burgievriendjes voor alle leuke reizen, kookavonden en weekends samen. Helena, Ellen, Marte, Ayla, Bob, Matthias, Ine, Sander, Holger, Jo, Stijn, Jasper, jullie zijn echte toppers!

Finally, I want to thank my family for their unwavering support. Un immense merci pour votre soutien et l'amour qui m'ont portée jusqu'ici. Sans vous, je ne serais pas la personne que je suis aujourd'hui. Sylvia et Isaac, merci d'être toujours là pour moi et d'être mes amis depuis toujours et pour toujours. Papa et maman, merci pour toutes les opportunités que vous nous avez offertes. Vous vous êtes souvent pliés en quatre pour nous permettre de nous épanouir, et je vous en suis éternellement reconnaissante. Yarni, Stéphanie, et les trois petits choux Johann, Sophie, et Théodore, je suis tellement heureuse que vous fassiez partie de la famille.

Thank you! Dankjewel! Merci!

Elisabeth Heremans

October 2024

Abstract

Sleep is a fundamental biological process essential for overall health and well-being. Studying it to diagnose sleep-wake disturbances requires overnight measurement of physiological signals, and annotation of these measurements through sleep staging. Traditionally, sleep studies are performed in the hospital using specialized polysomnography equipment and labor-intensive manual sleep staging. Wearable devices measuring brain activity and other physiological signals provide a more comfortable alternative and enable at-home sleep studies, but they necessitate automated sleep staging approaches.

This thesis focuses on the development and enhancement of automated sleep staging methods, which are crucial for home-based sleep studies and can reduce the workload in hospital-based sleep studies. The primary aim is to address challenges in performance, generalization, and trust in these methods to facilitate their adoption in clinical practice, and to validate the algorithms in clinical datasets. These objectives are attained in three main parts.

The first part explores techniques to improve the generalization of automated sleep staging methods trained on polysomnography, to achieve better performances on data from wearables. Two key methods are introduced: (1) supervised transfer learning via feature matching, which enhances performance using small amounts of labeled data by aligning features from different domains, and (2) adversarial domain adaptation, which minimizes domain mismatch by leveraging unlabeled data, and can be further improved with pseudo-labels and careful source domain selection. These methods demonstrate significant improvements in multiple clinical datasets.

The second part tackles the issue of trust and transparency in automated decision support systems by incorporating uncertainty estimation. The U-PASS framework is developed to optimize the reliability of sleep staging models through uncertainty-based training data selection and curation, active learning with clinician feedback, and deferral of uncertain predictions to experts. This

approach enhances both model performance and user trust, especially in clinical settings. The application of U-PASS to wearable sleep monitoring highlights the need to adjust handling of uncertainty in home-based wearable sleep monitoring applications.

The third part validates the developed algorithms across various clinical populations, demonstrating their effectiveness in characterizing sleep in patients with epilepsy, Alzheimer's disease, and Parkinson's disease. The results highlight that sleep staging can not only reproduce established clinical correlations but also contribute to therapeutic applications and aid in the diagnostic screening of brain disorders. Notably, sleep measurements using wearables are shown to be valuable in Alzheimer's disease screening. Depending on data availability, different transfer learning approaches are recommended: fine-tuning for polysomnography datasets and either feature matching or adversarial domain adaptation for wearable datasets.

In summary, this work contributes significantly to the integration of artificial intelligence in sleep medicine by addressing performance and generalization challenges, improving trust and transparency, and validating methods in clinical settings. The proposed solutions pave the way for more reliable and widespread use of automated sleep staging tools, potentially transforming both clinical diagnostics and large-scale sleep studies.

Beknopte samenvatting

Slaap is een fundamenteel biologisch proces, dat essentieel is voor menselijke gezondheid en welzijn. Het bestuderen ervan om slaap- en waakstoornissen te diagnosticeren vereist het 's nachts meten van fysiologische signalen en de annotatie van deze metingen door het classificeren van slaapstadia. Traditioneel worden slaapstudies uitgevoerd in het ziekenhuis, gebruik makend van polysomnografie met gespecialiseerde laboratoriumapparatuur en annotatie van deze metingen met arbeidsintensieve handmatige slaapstadiëring. Draagbare apparaten die de hersenactiviteit en andere fysiologische signalen meten, bieden een comfortabeler alternatief en maken slaapstudies in de thuisomgeving mogelijk, maar vereisen wel geautomatiseerde slaapstadiëring.

Deze thesis richt zich op de ontwikkeling en verbetering van geautomatiseerde methoden voor slaapstadiëring, die cruciaal zijn voor thuis-slaapstudies en de werklast bij slaapstudies in het ziekenhuis kunnen verminderen. Het doel is om uitdagingen op het gebied van performantie, generalisatie en vertrouwen in deze methoden aan te pakken om de adoptie van deze methoden in de klinische praktijk te vergemakkelijken, en om de algoritmen te valideren in klinische datasets. Deze doelstellingen worden bereikt in drie hoofdonderdelen.

Het eerste deel onderzoekt technieken om de generalisatie te verbeteren van geautomatiseerde methoden voor slaapstadiëring. Het doel is om modellen getraind op polysomnografie betere performanties te doen behalen op signalen van draagbare apparaten. Twee belangrijke methoden worden geïntroduceerd: (1) gesuperviseerde *transfer learning* via *feature matching*, die de performantie verbetert met behulp van kleine hoeveelheden geannoteerde data door *features* van verschillende types signalen met elkaar te aligneren, en (2) *adversarial domain adaptation*, die domeinverschillen minimaliseert door gebruik te maken van ongelabelde data, en verder kan worden verbeterd met pseudo-labels en zorgvuldige selectie van het oorspronkelijke domein. Deze methoden zorgen voor aanzienlijke verbeteringen in meerdere klinische datasets.

Het tweede deel pakt het probleem van vertrouwen en transparantie in geautomatiseerde beslissingsondersteunende systemen aan door schattingen van onzekerheid te introduceren. Het U-PASS-framework wordt ontwikkeld om de betrouwbaarheid van slaapstadiëringsmodellen te optimaliseren door het selecteren en samenstellen van *training data* op basis van onzekerheid, actief leren met feedback van clinici en doorschuiven van onzekere voorspellingen naar experts. Deze aanpak verbetert zowel de performantie van het model als het vertrouwen van de gebruiker, met name in klinische settings. De toepassing van U-PASS op draagbare slaapmonitoring benadrukt de noodzaak om het gebruik van onzekerheid aan te passen in thuisgebaseerde draagbare slaapmonitoringtoepassingen.

Het derde deel valideert de ontwikkelde algoritmen in verschillende klinische populaties. De effectiviteit van deze algoritmen wordt aangetoond in het karakteriseren van slaap bij patiënten met epilepsie, de ziekte van Alzheimer en de ziekte van Parkinson. De resultaten benadrukken dat slaapstadiëring niet alleen bestaande klinische correlaties kan reproduceren, maar ook kan bijdragen aan therapeutische toepassingen en diagnostische screening van hersenaandoeningen. Meer specifiek wordt aangetoond dat slaapmetingen met wearables waardevol zijn voor screening van de ziekte van Alzheimer. Afhankelijk van de beschikbaarheid van data worden verschillende *transfer learning*-benaderingen aanbevolen: *fine-tuning* is geschikt voor polysomnografie-datasets, en *feature matching* of *adversarial domain adaptation* voor wearable-datasets.

In conclusie draagt dit werk aanzienlijk bij aan de integratie van artificiële intelligentie in slaapgeneeskunde door uitdagingen op het gebied van performantie en generalisatie aan te pakken, vertrouwen en transparantie te verbeteren en methoden te valideren in klinische settings. De voorgestelde oplossingen banen de weg voor meer betrouwbare en wijdverbreide toepassing van geautomatiseerde methoden voor slaapstadiëring. Deze methoden bieden potentieel tot vernieuwingen binnen de klinische diagnostiek en voor grootschalige slaapstudies.

Contents

Abstract v

Beknopte samenvatting vii

Contents ix

List of Acronyms xv

List of Symbols xvii

List of Figures xix

List of Tables xxiii

I Introduction 1

1 Introduction 2

1.1 Sleep . . . 3
1.1.1 Sleep stages . . . 3
1.1.2 Sleep disturbances . . . 3
1.1.3 Sleep and Alzheimer's disease . . . 5
1.2 Studying sleep and sleep-wake disturbances . . . 5
1.2.1 Polysomnography . . . 5
1.2.2 Wearable monitoring . . . 6
1.2.3 Sleep scoring . . . 7
1.3 Automated sleep analysis . . . 8
1.3.1 Machine learning . . . 9
1.3.2 Deep learning . . . 11
1.3.3 Deep learning methods for sleep staging . . . 11
1.3.4 Performance metrics . . . 13

1.4 Scientific knowledge gaps to address . . . 14
1.5 Research objectives and overview . . . 16
1.5.1 Research objectives . . . 16
1.5.2 Chapter overview . . . 17

2 Materials 23
2.1 Datasets . . . 24
2.1.1 Healthy . . . 24
2.1.2 Diseased . . . 25
2.2 Sleep staging networks . . . 29
2.2.1 Pre-processing . . . 29
2.2.2 Attention-based recurrent neural network . . . 30
2.2.3 SeqSleepNet . . . 30
2.2.4 Parameters and settings . . . 31
2.2.5 Rationale behind the architecture choice . . . 32

II Transfer learning approaches to improve generalization 33

3 Feature matching as an improved transfer learning technique 35
3.1 Introduction . . . 36
3.2 Materials . . . 38
3.2.1 Source domain . . . 38
3.2.2 Target domains . . . 38
3.2.3 Sleep staging networks . . . 39
3.3 Feature matching . . . 39
3.3.1 Transfer learning framework . . . 39
3.3.2 Feature matching method . . . 40
3.4 Experiments . . . 42
3.4.1 Transfer learning scenarios . . . 42
3.4.2 Experimental setup . . . 42
3.4.3 Minibatch construction . . . 44
3.4.4 Training parameters . . . 44
3.5 Results . . . 45
3.5.1 Baseline performance on the source domain . . . 45
3.5.2 Fine-tuning and feature matching performance on the target domains . . . 45
3.5.3 Visualizing the feature spaces . . . 49
3.6 Discussion . . . 49
3.7 Conclusion . . . 53

4 Unsupervised and semi-supervised adversarial domain adaptation 55
4.1 Introduction . . . 56

4.2 Methods . . . 57
4.2.1 Domain adaptation . . . 57
4.2.2 Adversarial domain adaptation . . . 58
4.2.3 Basic method . . . 59
4.2.4 Pseudo-labels . . . 60
4.2.5 Target labels . . . 61
4.3 Materials . . . 62
4.3.1 Source domain: PSG datasets . . . 62
4.3.2 Target domain: wearable EEG datasets . . . 62
4.3.3 Sleep staging network . . . 63
4.4 Experiments . . . 63
4.4.1 Experimental setup . . . 63
4.4.2 Influence of target (pseudo-)labels . . . 64
4.4.3 Influence of amount of domain mismatch . . . 65
4.4.4 Personalization . . . 66
4.5 Results . . . 67
4.5.1 Influence of target (pseudo-)labels . . . 67
4.5.2 Influence of domain mismatch . . . 68
4.5.3 Personalization . . . 71
4.6 Discussion . . . 72
4.7 Conclusion . . . 74

III Uncertainty estimation to increase trust 75

5 An uncertainty-guided deep learning pipeline for automated sleep staging 77
5.1 Introduction . . . 78
5.2 Methods . . . 81
5.2.1 Data and study cohorts . . . 81
5.2.2 Model and training . . . 81
5.2.3 Formulation of uncertainty . . . 81
5.2.4 U-PASS pipeline . . . 82
5.3 Results . . . 86
5.3.1 Insights from uncertainty . . . 86
5.3.2 Data collection . . . 86
5.3.3 Data selection . . . 87
5.3.4 Active learning . . . 89
5.3.5 Deferring uncertain test samples . . . 90
5.4 Discussion . . . 92
5.5 Conclusion . . . 95
Appendices . . . 96
5.A Extended deferral results . . . 96

5.B U-PASS in Dreem Open Datasets . . . 96
5.C Inter-rater variability . . . 97

6 Automated remote sleep monitoring needs uncertainty quantification 101
6.1 Introduction . . . 102
6.2 Methods . . . 103
6.2.1 Dataset and sleep staging model . . . 103
6.2.2 U-PASS and uncertainty measures . . . 104
6.2.3 Manual review . . . 105
6.2.4 Reliability assessment . . . 106
6.3 Results . . . 106
6.3.1 Uncertainty-based training data curation . . . 106
6.3.2 Detecting uncertainty at deployment time . . . 107
6.3.3 Does manual review help for uncertain data? . . . 109
6.3.4 Is automated sleep staging on wearable EEG reliable enough? . . . 111
6.4 Discussion . . . 112
6.5 Conclusion . . . 119
Appendices . . . 120
6.A Experimental setup . . . 120
6.B Sources of uncertainty . . . 121
6.C Bland-Altman plots . . . 121
6.D Sleep parameters' correlations with AHI . . . 122

IV Validation in clinical populations 125

7 Automated sleep staging in clinical practice 127
7.1 Introduction . . . 129
7.2 Parkinson's disease . . . 131
7.2.1 Dataset . . . 131
7.2.2 Methods . . . 131
7.2.3 Results . . . 132
7.3 Epilepsy in pediatric population . . . 132
7.3.1 Dataset . . . 132
7.3.2 Methods . . . 132
7.3.3 Results . . . 133
7.4 Epilepsy in adult population . . . 134
7.4.1 Dataset . . . 134
7.4.2 Methods . . . 135
7.4.3 Results . . . 135
7.5 Alzheimer's disease . . . 136
7.5.1 Dataset . . . 136

7.5.2 Methods . . . 138
7.5.3 Results . . . 139
7.6 Discussion . . . 139
7.7 Conclusion . . . 142

8 Wearable sleep recordings for Alzheimer's disease screening 145
8.1 Introduction . . . 146
8.2 Methods . . . 148
8.2.1 Datasets and PSG scoring . . . 148
8.2.2 Sleep staging model . . . 150
8.2.3 Feature calculation . . . 150
8.2.4 Detection of patients with AD . . . 150
8.3 Results . . . 151
8.3.1 Sleep staging results . . . 151
8.3.2 Detection of patients with AD . . . 151
8.3.3 Insights into the hypnogram and physiological features . . . 153
8.4 Discussion . . . 155
8.4.1 AD detection performance . . . 156
8.4.2 Sleep features related to AD . . . 157
8.4.3 Automatic sleep scoring . . . 158
8.4.4 Implications for future research . . . 158
8.4.5 Limitations and strenghts . . . 159
8.5 Conclusion . . . 159
Appendices . . . 160
8.A Supplementary methods . . . 160
8.B Supplementary results . . . 164

V Conclusion 183

9 Conclusion 185
9.1 Main findings . . . 185
9.1.1 Part II - Transfer learning approaches for wearable EEG data . . . 185
9.1.2 Part III - Uncertainty estimation to increase trust . . . 186
9.1.3 Part IV - Validation in clinical populations . . . 188
9.2 Future research directions . . . 189
9.2.1 Open technical challenges in automated sleep staging . . . 189
9.2.2 Future perspectives: AI propelling sleep research forward 191
9.3 Final thoughts . . . 193

Appendices **196**

A Explaining uncertainty in AI **197**
A.1 Introduction . . . 198
A.2 Uncertainty explanations . . . 198
A.3 Experimental methodology . . . 199
A.3.1 Synthetic dataset: two moons . . . 199
A.3.2 MNIST dataset . . . 201
A.4 Results . . . 203
A.4.1 Synthetic dataset: two moons . . . 203
A.4.2 MNIST dataset . . . 205
A.5 Discussion . . . 209

Bibliography **211**

Acknowledgments **235**

Curriculum vitae **237**

List of Publications **239**

List of Acronyms

A	**AASM**	American Academy of Sleep Medicine
	Aβ	Amyloid-beta
	ACM	Accelerometry
	AD	Alzheimer's disease
	AHI	Apnea-hypopnea index
	AI	Artificial intelligence
	ARNN	Attention-based recurrent neural network
	AUC	Area under the curve
B	**biRNN**	Bidirectional recurrent neural network
	BMI	Body mass index
C	**CDR**	Clinical Dementia Rating
	CIE	Cognitively intact elderly
	CNN	Convolutional neural network
E	**ECG**	Electrocardiography
	EEG	Electroencephalography
	EMG	Electromyography
	EOG	Electrooculography
	ESS	Epworth Sleepiness Scale
F	**FIR**	Finite impulse response
G	**GCN**	Graph convolutional network
	GRU	Gated recurrent unit
K	**KL**	Kullback-Leibler
	κ	Cohen's kappa score

M	**MF1**	Macro F1-score
	MLP	Multi-layer perceptron
	MMD	Maximum mean discrepancy
	MMSE	Mini-Mental State Examination
	MSE	Mean squared error
N	**N1**	Non-rapid eye movement sleep, stage 1
	N2	Non-rapid eye movement sleep, stage 2
	N3	Non-rapid eye movement sleep, stage 3
O	**OSA**	Obstructive sleep apnea
P	**PD**	Parkinson's disease
	PSG	Polysomnography
	PSQI	Pittsburg Sleep Quality Index
R	**RNN**	Recurrent neural network
	REM/R	Rapid eye movement sleep
	ROC	Receiver operating characteristic
W	**wF1**	Weighted F1-score

List of Symbols

General

$\mathcal{L}$ or L	Loss function
$\mathbb{R}$, $\mathbb{R}^m$, $\mathbb{R}^{m\times n}$	Set of real scalars, vectors of size m and matrices of size $m \times n$, respectively
$\mathbb{N}$	Set of non-negative integers
log	Logarithm
$p(y\|x)$	Conditional probability of y given x
$\|\cdot\|$	ℓ_2-norm
$\min_{X\in\mathcal{S}} f(X)$	Minimal value of function f over the constraint $X \in \mathcal{S}$
$\mathbb{E}[\cdot]$	Expectation operator

Frequently used notations

M	Sequence length: amount of epochs in a sequence
n	Amount of data points or recordings
$\mathcal{X}$	Input space
x	Input feature or data (often epoch in sleep staging)
$\mathcal{Y}$	Output space
y	Label of a data point (often a sleep stage in sleep staging)
$\hat{y}$	Predicted label by a model
p	Class probability output of a model
$f(\cdot)$	Predictive function / objective function
$P(\cdot)$	Marginal probability distribution
H	Entropy metric

List of Figures

1.1 Sleep is categorized into four stages. 4
1.2 Polysomnography exam. 6
1.3 The full pipeline of automated sleep staging, from capturing the physiological signals from a human to obtaining the corresponding hypnogram. 8
1.4 The sequence-to-sequence framework for sleep staging. 13
1.5 A schematic overview of the thesis. 21

2.1 Main setup for the EEG measuring wearable device SensorDot in this thesis. 26
2.2 Illustration of the two sleep staging networks used throughout this thesis. 31

3.1 The conceptual difference between fine-tuning and feature matching. 38
3.2 Feature matching consists of concurrently training two networks, while matching the feature representations. 41
3.3 Illustration of a minibatch for training the feature matching approach. 45
3.4 Absolute difference in accuracy of feature matching, fine-tuning and training from scratch with respect to direct transfer. 49
3.5 UMAP visualization of the C4-A1 and cross-head behind-the-ear modality features (both of the Leuven - SDS dataset), learned by SeqSleepNet, before and after transfer learning on 5 recordings of the Leuven - SDS dataset. 50

4.1 Schematic illustration of adversarial domain adaptation. 59
4.2 The accuracy of the three adversarial domain adaptation techniques on the three target datasets. 68
4.3 Sensitivity analysis to the hyperparameters λ_{adv} and λ_{ps} in Surrey - cEEGrid dataset. 70

4.4 The influence of the domain mismatch between the source domain and target domain. 70

5.1 Full uncertainty-guided pipeline for sleep staging (U-PASS). . . 80
5.2 (a) The average training dynamics for the 1% most data-ambiguous, model-ambiguous, easy-to-classify and hard-to-classify samples. (b) UMAP visualization of the feature distribution resulting from training a sleep staging network. . . 87
5.3 U-PASS applied to hospital-based PSG data. The resulting accuracy after every consecutive step is indicated with a red box. 88
5.4 Parameters causing data uncertainty in the PSG training dataset. 89
5.5 Hypnograms (sleep stages over time) for snippets of two different recordings, demonstrating how personalization through active learning (AL) improves the performance. 90
5.6 The accuracy on the test data depends on both the distance from the training data and the uncertainty of those training data. . . 92
5.A Extended results showing the performance of the different measures of uncertainty for deferring uncertain test samples. . 96
5.B U-PASS applied to the Dreem Open Datasets (DOD). 99
5.C The relation between the data uncertainty ('aleatoric uncertainty') and the inter-scorer variability for 28 randomly selected recordings used for training U-PASS on the DOD-O dataset. 100

6.1 The proposed machine learning pipeline for automated sleep staging on wearable sleep recordings, enhanced with uncertainty quantification. 103
6.2 U-PASS applied to sleep data captured by a wearable device. . 108
6.3 Comparison of automated sleep scoring to manual scoring for behind-the-ear wearable EEG recordings. 110
6.4 Clinical significance of sleep parameters derived from model predictions and PSG reference. 114
6.A Clinical parameters lie at the source of data uncertainty in the wearable training dataset. 121
6.B Bland-Altman plots for a number of important sleep parameters, where each sample is one patient. 122

7.1 Box-plot comparing REM sleep time according to ground truth scoring and scoring by the fine-tuned algorithm. 134
7.2 Representation of the DREEM2 device and the experimental procedure. 138
7.3 Normalized power in the slow wave frequency range of the 4-second window following the first stimulation/sham, the second stimulation/sham, and the total stimulations/shams. 140

8.1 Overview of the methodology. 148
8.2 The performance in detecting patients with AD and the subset with prodromal AD. 152
8.3 Correlation between MMSE and sleep features based on wearable data. 154
8.4 Most selected wearable-based and PSG-based physiological features for the AD detection model. 155
8.A Comparison of the sleep parameters of different patient subgroups based on ground truth, PSG-based AI and wearable-based AI. . 173
8.B ROC performance curves for detecting AD with and without OSA. 176
8.C ROC performance curves for detecting AD in younger and older subjects. 177
8.D ROC performance curves for discriminating AD from CIE from each of the two datasets. 179
8.E ROC performance curves for discriminating AD and prodromal AD from CIE using physiological features and using clinical data from questionnaires. 180
8.F Correlation between sleep features based on the PSG data and MMSE. 181
8.G Most selected physiological features based on ground truth scoring and AI scoring. 182

A.1 The synthetic two moons dataset, which forms a binary classification problem. 200
A.2 Digits from the MNIST dataset were altered by striking them through, to create out-of-distribution data. 202
A.3 Dependence plots, each showing the effect of a single feature on the estimated epistemic uncertainty. 203
A.4 Epistemic uncertainty visualizations. 204
A.5 Dependence plots, each showing the effect of a single feature on the estimated aleatoric uncertainty. 205
A.6 Aleatoric uncertainty visualizations. 206
A.7 A line that strikes through the MNIST dataset's digits is detected as a source of epistemic uncertainty. 207
A.8 The MNIST dataset and its sources of epistemic uncertainty visualized. 208
A.9 UMAP visualization showing the output of the last hidden layer of the neural network architecture trained on the MNIST dataset. 209
A.10 Explanations for five of the most common confusions between digits. 210

List of Tables

2.1 Summary of the datasets used in this thesis. 28

3.1 The datasets and their subdivision into training, validation and test sets for every cross-validation round. 43
3.2 The distribution of the different classes in the three datasets. . 43
3.3 Sleep staging performance of ARNN and SeqSleepNet on source and target domains, using feature matching, fine-tuning, fine-tuning with KL-divergence regularization, direct transfer and training from scratch. 47
3.3 Sleep staging performance of ARNN and SeqSleepNet on source and target domains - continued 48

4.1 The three different adversarial domain adaptation methods. . . . 61
4.2 The results of the three adversarial domain adaptation techniques: the basic version using no labels of the target domain, the pseudo-labels version and the version with target domain labels of two recordings. 69
4.3 Results of personalization on the three target domains. 71

6.1 Sample characteristics. 104
6.2 A total of 26 sleep parameters are calculated based on the hypnograms obtained with our deep neural network and the reference hypnograms, as annotated by a clinician based on PSG. Pearson's correlation coefficient (ρ) is calculated between the sleep parameters calculated with both approaches, and the t-test for related samples is performed to identify significant differences between both distributions. 113

6.A A total of 26 sleep parameters calculated based on the hypnograms obtained with our deep neural network and the reference hypnograms, as annotated by a clinician based on PSG. Pearson's correlation coefficient (ρ) is calculated between the sleep parameters and the AHI on total sleep time (TST). . . . 123

7.1 Summary of the datasets used for clinical validation in this chapter. 130
7.2 Transfer learning approaches and the ideal scenarios in which to use them. . . . 131
7.3 Sleep staging performance of SeqSleepNet on pediatric epilepsy population. . . . 133
7.4 Sleep staging performance of SeqSleepNet on adult epilepsy population. . . . 136
7.5 Sleep parameters estimated from the wearable and from the 25-channel EEG, and Spearman correlation between both sets of parameters. . . . 137
7.6 Sleep macrostructure for nights 2 and 16. The results as shown represent respectively those by the embedded automatic algorithm of the DREEM2 headband and those of the scoring of SeqSleepNet with expert verification. . . . 139

8.1 Sample characteristics of the used portion of the Leuven – SensorDot Senior (SDS) dataset . . . 149
8.2 Sample characteristics of the Leuven – SensorDot AD (SDAD) dataset . . . 149
8.3 Sample characteristics of the test dataset . . . 150
8.A Hypnogram features and their meaning. . . . 162
8.B Biomarkers for the AD patients of the Leuven - SensorDot Alzheimer's disease (SDAD) dataset. . . . 165
8.B Clinical biomarkers for the AD patients of the Leuven - SensorDot Alzheimer's disease (SDAD) dataset - continued . . . 166
8.B Clinical biomarkers for the AD patients of the Leuven - SensorDot Alzheimer's disease (SDAD) dataset - continued . . . 167
8.B Clinical biomarkers for the AD patients of the Leuven - SensorDot Alzheimer's disease (SDAD) dataset - continued . . . 168
8.B Clinical biomarkers for the AD patients of the Leuven - SensorDot Alzheimer's disease (SDAD) dataset - continued . . . 169
8.B Clinical biomarkers for the AD patients of the Leuven - SensorDot Alzheimer's disease (SDAD) dataset - continued . . . 170
8.B Clinical biomarkers for the AD patients of the Leuven - SensorDot Alzheimer's disease (SDAD) dataset - continued . . . 171
8.C Sleep staging performance for PSG and wearable data, compared between pre-trained and personalized sleep staging models. . . . 172

8.D Performances in detecting patients with AD and the subgroup in the prodromal AD stage based on hypnogram features and physiological features. 175

A.1 CNN architecture used for MNIST dataset 201

Part I

Introduction

1 Introduction

We spend one third of our lives sleeping, and yet, this complex biological process taking place in our brain still holds many mysteries. Why do we sleep? Is sleep composed of different 'stages', or is it a continuous process? What is 'good' or 'bad' sleep, and how does it affect us? Although not all of sleep's functions are entirely understood, numerous studies have shown that the importance of good sleep can hardly be overstated. It has powerful effects on the immune system [1], impacts brain function and behavior [2, 3], restores energy, and helps the body repair and regenerate [4]. With sleep fulfilling so many primary functions, it comes as no surprise that long-term sleep disturbances are linked to a myriad of medical conditions, from obesity and diabetes to brain disorders [5, 6].

To better understand sleep and diagnose its associated disorders, we need to accurately measure and characterize it, but this is precisely where modern medicine still falls short. In sleep medicine, the current standard practice for measuring sleep is an overnight recording of multiple physiological signals in the hospital, using bulky hospital equipment. Trained clinicians visually analyze and score the measured signals to quantify sleep and diagnose sleep disorders. This form of sleep measurement is very precise, but also costly, impractical and uncomfortable for the patient. As a result, it is mostly employed for diagnosis of sleep-related breathing disorders, and most sleep problems still go undiagnosed and untreated.

With the upcoming of wearable devices allowing cheap, comfortable monitoring at home, it will become feasible to perform long-term sleep monitoring on the general population, and to better understand the complex relations between sleep problems and associated diseases. These wearable devices call for reliable data processing algorithms that automatically extract sleep stages and relevant sleep characteristics.

1.1 Sleep

1.1.1 Sleep stages

Sleep is not a uniform brain state, but instead, a complex set of processes that fulfill several physiological functions. According to current clinical practice, it is categorized into four different stages. With wake as a fifth stage, we speak of five distinct vigilance states: wake (W), rapid eye movement (REM) sleep, and non-REM sleep stages N1, N2, and N3 [7]. Each vigilance stage encompasses variations in brain activity, eye movement, and muscle tone. When we think of a good night's sleep, we often focus on the total sleep duration. However, this is not the only important factor in determining how rested you are. Sleep quality plays a critical role in determining how restorative the time spent sleeping was. A healthy night of sleep is comprised of multiple sleep cycles, in which you progress through the different sleep stages (Figure 1.1). The sleep cycle plays a vital role in many sleep disorders, like obstructive sleep apnea (OSA) and insomnia. Typically, a night of sleep should be comprised of four to six sleep cycles, averaging 90 minutes per cycle [8]. Both the length of these cycles and their composition - how much time is spent in each stage - varies throughout the night. Sleep cycles are also highly variable between individuals and from night to night, based on a number of factors such as caffeine and alcohol intake, age, and disease [6, 9].

1.1.2 Sleep disturbances

Sleep is crucial to a person's wellbeing and mental, emotional and physical health. When changes in sleep (both in the cyclicity, patterns, or duration) lead to daytime impairment, we speak of sleep-wake disturbances. Sleep-wake disturbances have a hugely negative impact on quality of life and longetivity, and affect an estimated 20-30% of the adult population [11]. 29.8% of adults in the United States have trouble sleeping, while 27.2% suffer from daytime sleepiness [12]. These sleep-wake disturbances have a large toll on global health and society, causing a higher mortality and lower productivity. As a result, sleep problems are estimated to cost around 2% of the GDP of developed countries, amounting to a worldwide yearly cost of some 700 billion USD [13].

There are numerous types of sleep-wake disturbances, with around 90 recognized sleep disorders [11]. Most are characterized by at least one of the following

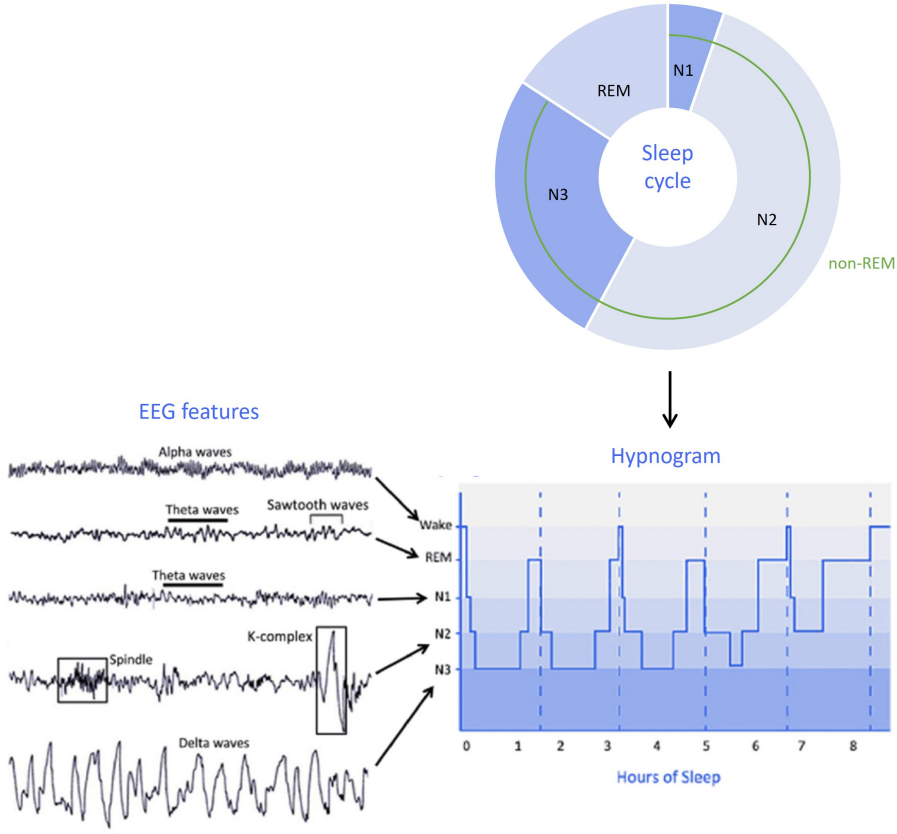


Figure 1.1: Sleep is categorized into four stages. Throughout the night, individuals progress through multiple sleep cycles, sequentially transitioning through these sleep stages. Sleep stages are recognized by looking at characteristic patterns and frequency content of physiological signals, mainly brain activity (electroencephalography or EEG). In five-stage sleep scoring, we are counting the wake state as well, which is technically not a sleep stage, but a vigilance state. The sleep architecture or hypnogram is obtained by annotating each 30-second segment of the measured signals as a vigilance state. (Image partly from [10].)

symptoms: difficulty to maintain or initiate sleep, excessive daytime sleepiness, and abnormal events during sleep. A few of the most common sleep-wake disturbances are insomnia, with 33%-50% of adults experiencing insomnia symptoms [14], and OSA, with a prevalence of about 20% [15].

1.1.3 Sleep and Alzheimer's disease

As sleep is critical for cognitive processing and memory consolidation, it comes as no surprise that sleep disturbances are associated with cognitive impairments. Sleep disturbances are very common in dementia [11], of which Alzheimer's disease (AD) is the leading cause. As such, poor sleep is one of the most prevalent and disruptive symptoms for patients with AD and among the earliest symptoms in preclinical AD [16,17]. Approximately two-thirds of AD patients have abnormal sleep patterns [18, 19] and there is plenty of evidence for a complex interplay between sleep and AD [17,20]. On the one hand, AD causes sleep-wake disturbances [21]. On the other hand, sleep-wake disturbances cause an increase of amyloid-beta (Aβ), the main component of the plaques that AD patients form in their brain [22], accelerating AD progression [23]. Sleep disorders affect behavior, caregiving and cognition, and are thereby the primary reason for institutionalization of AD patients [24,25]. Therefore, they play an important role in the heavy burden of AD on caregivers, the healthcare system, and its financing.

Although associations between sleep and AD have been reported, the precise interactions between them are not fully understood yet. Are sleep-wake disturbances a mere result of AD, or do they drive its gradual progression? Despite their central role in the disease, sleep-wake disturbances are heavily understudied in AD. The reason for this is the lack of appropriate monitoring technology that can be used to investigate interactions over longer periods of time in real-life, and capture the night-to-night variability in sleep and disease symptoms.

1.2 Studying sleep and sleep-wake disturbances

1.2.1 Polysomnography

The gold standard for studying sleep and diagnosing sleep-related disorders is currently based on full night polysomnography (PSG). This procedure, mostly performed in-hospital, entails recording brain activity with electroencephalography (EEG) in combination with other electrophysiological signals such as

heart rhythm through electrocardiography (ECG), eye movements through electro-oculography (EOG), and muscle activity of the chin and legs through electromyography (EMG) [26] (Figure 1.2). Body position and breathing functions are also measured.

PSG recordings are usually performed in a laboratory setting, because expertise is required to apply the EEG electrodes. However, these laboratory measurements are associated with a high cost (1300 euros per night in Belgium [27]), limited availability, and discomfort for the patient. Therefore, the utility of PSG is limited to very specific disorders (e.g. sleep apnea), and not suited for longitudinal sleep monitoring or screening in the general population, including AD patients.

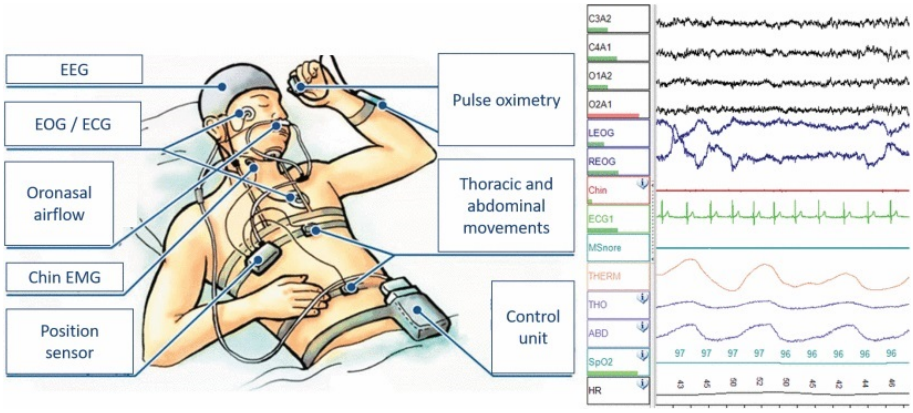


Figure 1.2: A representation of a PSG examination and example traces of the signals obtained from the different sensors. (Image: [28].)

1.2.2 Wearable monitoring

The recent surge of wearable devices presents an alternative to PSG with potential for prolonged, home-based sleep monitoring on a population level. Indeed, the abundance of consumer-grade devices enables anyone to measure their sleep, offering several advantages over traditional PSG. First, consumer-grade wearable devices are more accessible and affordable than standard PSG equipment. Second, their more minimalistic setup generally makes them more comfortable than PSG, which is especially important for long-term monitoring. However, this minimal setup also has drawbacks, like the risk of collecting incomplete information. Additionally, wearables sometimes suffer from lower

signal quality, which can further degrade if users improperly wear the devices or move during sleep.

Wearables typically collect one or more sensing modalities, with EEG, ECG, accelerometry (ACM), and photoplethysmography being the most frequently used signals [29]. Among these, EEG is the most reliable for five-stage sleep scoring [29]. This is unsurprising, as the stages of sleep (see Figure 1.1) are defined based on the interpretation of EEG signals. While PPG- and ACM-based wearables measure other physiological signals correlated with sleep stages, they do not allow for perfect discrimination of these stages. Therefore, when fine-grained sleep analysis is required, particularly for medical use and clinical diagnosis, EEG remains the gold standard for sleep measurement.

Miniaturized wearable brain recording devices typically have EEG electrodes in non-standardized positions that improve user comfort, resulting in new EEG derivations like behind-the-ear-EEG [30,31], in-ear-EEG [32,33], and forehead EEG [34]. Several previous studies have simultaneously collected high-quality standard EEG data and wearable EEG data in laboratory settings. These studies have demonstrated that it is feasible to extract relevant sleep information from data recorded from non-conventional locations [32,35,36].

1.2.3 Sleep scoring

To gather meaningful information from signals recorded by PSG or wearable devices, the data must be analyzed, and relevant patterns or events need to be identified and labeled. In current clinical practice, PSG recordings are visually reviewed and manually annotated by a certified scorer to identify relevant sleep characteristics, according to American Academy of Sleep Medicine (AASM) scoring guidelines [7]. Sleep scoring primarily involves *sleep staging*, a procedure in which a sleep stage is assigned to every thirty-second epoch to form a sequence of sleep stages known as a *hypnogram* (Figure 1.1).

Sleep stages The AASM guidelines define five sleep stages based on the identification of specific oscillatory patterns found in PSG data (see Figure 1.1). W is characterized by high-frequency, low-amplitude waves in the EEG. Non-REM sleep is marked by increasing amplitude and decreasing frequency in EEG signals and is divided into three stages: N1, N2, and N3. N1 shows nearly no low-frequency, high-amplitude activity, N2 exhibits a bit more, and N3 is associated with pronounced low-frequency, high-amplitude activity, which is why it is sometimes referred to as *slow-wave sleep*. REM sleep looks similar to W and N1 based on EEG data, although it also exhibits sharp transient waves

called *sawtooth waves*. Rapid eye movements observed in EOG and low muscle tone observed in EMG signals further help distinguish REM from the other stages [37].

In addition to annotating sleep stages, sleep scoring also includes identifying cardiac events, breathing-related events like apneas and hypoventilation, movements, and arousals. Key sleep characteristics, such as the sequence and duration of sleep stages, along with the timing and length of other events, are extracted from the annotated sleep recordings. These, in turn, are used to diagnose sleep-wake disturbances.

1.3 Automated sleep analysis

The manual scoring of PSG can take up to two hours and is subject to substantial inter-rater and intra-rater variability. To help alleviate these issues, researchers have been investigating the use of automatic aids for sleep scoring since many years. The rise of unobtrusive and user-friendly wearable at-home monitoring devices has made the need for automated interpretation even more pressing, as these wearables allow to capture large amounts of sleep data to screen for or track sleep-wake disturbances. Manually annotating large amounts of longitudinal sleep data would be extremely time-consuming, costly, and overall unfeasible. Most importantly, it has been shown that trained electrophysiologists are unable to accurately annotate data from novel wearable devices [35], as AASM scoring rules are only defined for standardized recording positions on the head and would need to be adapted for analyzing data recorded from new electrode configurations. Addressing these challenges requires machine learning approaches that can automate sleep scoring (Figure 1.3).

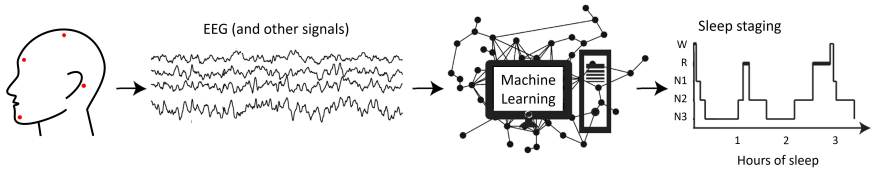


Figure 1.3: The full pipeline of automated sleep staging, from capturing the physiological signals from a human to obtaining the corresponding hypnogram.

1.3.1 Machine learning

This section introduces a few fundamental concepts of machine learning. Essentially, machine learning means teaching a computer to make predictions by providing it with learning examples. In this case, we are interested in machine learning methods that train a computer to predict sleep stages when fed with physiological signals. The parametric function built to achieve these predictions is called a predictor, model, or algorithm. Building this function requires a process called *learning* or *training*, which can be achieved with different strategies or *learning paradigms*. Machine learning predictors can be categorized into *shallow learning* and *deep learning* predictors.

Learning or training

The operations performed by a machine learning predictor are not fixed; instead, they depend on the model's inner parameters, which determine how inputs are processed into outputs. To achieve desired predictions, these parameters are trained to optimize a certain objective function that depends on the specific task the model needs to achieve. The objective function quantifies the quality of the model's output when processing a sample.

During training, samples are passed through the model, the objective function measures the quality of the output, and optimizers compute updates which are applied to the parameters to improve performance. This general procedure applies to various types of machine learning models. For a specific type of machine learning models called *neural networks*, this training process involves backpropagation, where the objective function's feedback is propagated through the model to adjust the parameters.

Learning paradigms

- **Supervised learning**

  Most commonly, machine learning models are trained in a supervised fashion, which means that they learn from labeled data, i.e. data (x) for which the true target output (y) is known. Supervised learning then refers to the process in which the network learns the mapping of an input to an output. The objective function that is optimized typically quantifies the error between the predicted output and the true label. This type of learning leads to concrete predictions, but it requires training datasets consisting of large amounts of labeled data.

- **Unsupervised learning**

  Unsupervised learning means learning from data without any labels. Since labeled data can be expensive to acquire, especially in biomedical applications, this can be an alternative to train machine learning methods on large amounts of unlabeled data when labeled data is not available. However, as the desired outputs are not known by the neural network, the use of unsupervised learning is limited to abstract tasks like clustering the data or reducing its dimensionality.

- **Semi-supervised learning**

  Trying to combine the advantages of supervised learning (learning a concrete prediction task) and unsupervised learning (learning from inexpensive, unlabeled data), semi-supervised learning combines both approaches. This can be advantageous when few labeled data and large amounts of unlabeled data are available, which is often the case in biomedical applications.

Shallow and deep learning models

Machine learning algorithms broadly fall into two categories: shallow learning methods that learn from manually extracted features and deep learning methods that directly learn from raw data. The shallow learning algorithms rely on engineered feature extraction, based on prior knowledge about the data and the prediction task. These features, specifically designed to be able to discriminate between the classes, are fed to a classifier, which learns the optimal decision boundary based on the set of features and their corresponding classes.

Deep learning methods, on the other hand, are characterized by their ability to learn directly from raw data. These models do not rely on prior knowledge and explicitly hand-crafted features, but learn the optimal features while they are being trained for classification. An important advantage of deep learning methods over feature-based approaches is that they can efficiently exploit large quantities of data. As a result, deep learning has become the dominant machine learning approach for various tasks. It has seen numerous successes across many domains including computer vision, natural language processing, and speech recognition. The advantages of deep learning come with an important downside: its accuracy gains come at the cost of interpretability. As deep learning methods consist of multi-layer networks which model highly non-linear functions, it is humanly impossible to fully understand them.

1.3.2 Deep learning

Deep learning algorithms are based on a class of machine learning models called neural networks. The term 'deep' refers to the fact that they are composed of multiple layers. Deep neural networks are essentially composed of layers of simple, differentiable operations like matrix multiplications and convolutions, which are in turn composed of multiplications and summations. Between those layers, nonlinear activation functions like Rectified Linear Units (ReLUs) [38] are inserted to allow for modeling non-linear functions.

Deep learning models can be designed into different architectures, each tailored to handle different types of data and tasks. The standard type is the multi-layer perceptron (MLP), in which the output of each layer is computed through matrix multiplication between the input data and a weight matrix. This architecture is typically used for tasks where data can be represented in a flat, tabular form.

In convolutional neural networks (CNNs), the operations that make up the network's layer are convolutional filters. As samples are passed through the network, these filters extract increasingly complex features, making CNNs particularly effective for image and spatial data processing.

Recurrent neural networks (RNNs) are designed for sequential data, such as time series. They have built-in memory through feedback loops that feed outputs back into the network as inputs. This architecture allows RNNs to consider both the current input and previous inputs, making them well-suited for processing sequential data.

Transformer architectures, a more recent development, rely on self-attention mechanisms to learn relationships between inputs at different positions. Unlike RNNs, transformers process sequential data by considering the entire sequence simultaneously, enabling them to capture long-range dependencies more effectively. This makes transformers highly suitable for tasks involving complex sequential data, such as natural language processing and machine translation.

1.3.3 Deep learning methods for sleep staging

From a technical machine learning perspective, sleep staging is an imbalanced multi-class classification task. The imbalanced nature of the prediction task comes from the fact that the five sleep stages are not equally represented in the data, with N1 being much scarcer than N2. Multi-class refers to the fact that we have to distinguish between more than two, in this case between five, classes. Machine learning tasks are typically divided into regression and classification

problems. In regression problems, the predicted output is a continuous variable, whereas in classification problems, it is a categorical output.

Over the past few decades, various machine learning approaches have been investigated for automated sleep staging, as highlighted by several review papers [39–42]. Shallow learning methods, however, did not achieve widespread adoption in clinical practice due to their unsatisfactory accuracy. More recently, deep learning began to be explored for EEG and PSG processing, gaining attention due to its successes in computer vision and language processing.

As public sleep datasets have grown in size, deep neural networks have demonstrated the ability to efficiently leverage large quantities of data. A recent review of deep learning approaches for sleep staging has demonstrated how the transition from shallow to deep learning has significantly improved sleep staging performance [39]. With the rising popularity of deep learning, numerous deep learning methods for automated sleep staging have been developed in recent years. Commonly used architectures include CNNs [43–45], RNNs [46–48], and more recently, transformer-based architectures [49, 50]. In most studies, neural networks were applied either directly to raw PSG data or to precomputed spectrograms, which are spectral images that represent the signals' frequency content over time.

CNNs, RNNs, and transformer-based neural networks are all capable of learning complex information from time series data, such as PSG and EEG. However, even these clever neural network architectures proved insufficient to achieve expert-level accuracy in sleep staging when naively applied to score one epoch at a time [39]. Sleep is governed by gradually transitioning physiological processes [51], and therefore, sleep experts manually score sleep stages by considering a larger context around each target epoch [7]. To take into account these temporal dynamics and reach expert-level sleep staging accuracy, deep learning methods have to learn the relevant features within each 30-second epoch, as well as model the long-range dependencies. In technical terms, this can be achieved by treating sleep staging as a time series classification problem on two levels: the epoch-level and the sequence-level.

Phan et al. [46] initially formulated this idea and introduced the sequence-to-sequence framework for sleep staging to model and learn these inter-epoch dependencies (see Figure 1.4). This framework processes a sequence of inputs (epochs) and generates a corresponding sequence of outputs (sleep stages), thus incorporating contextual information from both the sequence of epochs and the resulting sequence of sleep stages. It typically consists of an epoch encoder and a sequence encoder [41, 46]. The epoch encoder learns to extract relevant features from individual epochs, and the sequence encoder encodes the interaction between different epochs within a sequence. The epoch encoder

and sequence encoder are jointly optimized in an end-to-end fashion, so as to allow the two to interact. The sequence-to-sequence framework has become the dominant approach for sleep staging, largely outperforming previously used frameworks [43, 46, 47, 50]. Both the epoch encoder and sequence encoder can be implemented as any of the previously mentioned architectures which are appropriate for handling time series. Various studies introduced different implementations with a variety of combinations of architectures in the epoch and sequence encoder [43, 46–48, 50, 52–56]: CNNs, RNNs, transformers, and other architectures including graph convolutional networks (GCNs) [53] and hybrid networks (like CRNNs) [55].

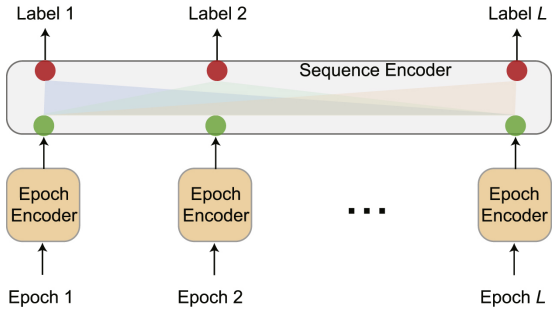


Figure 1.4: The sequence-to-sequence framework for sleep staging. The shaded regions in the sequence processing block illustrate that different epochs in the input sequence are affected by different contexts, which depend on each epoch's position in the sequence. The epoch encoder functions as an epoch-wise feature extractor, transforming each epoch into a feature vector representation (green circle). The sequence encoder enhances this representation (red circle), by incorporating interactions between each epoch and other epochs within its context. Image: [41].

1.3.4 Performance metrics

Various performance metrics are employed to compare the performance of different sleep staging methods presented in the literature. Some of the most common ones are the accuracy, Cohen's kappa score (κ), the macro F1-score (MF1), the weighted F1-score (wF1), and the per-class sensitivity and specificity. κ measures the inter-rater agreement between the scorer and the model. F1-score is the harmonic mean of recall and precision for a binary classification problem. In multi-class classification problems like sleep staging, MF1 is the unweighted mean of the per-class F1-scores, and wF1 is the mean weighted by each class's number of true instances.

In general, machine learning methods strive for 100% accuracy. However, in the case of sleep staging, this goal is unattainable. There is considerable disagreement between sleep experts on how to correctly score sleep, making it a subjective task with significant inter-rater disagreement. The main reason for this ambiguity lies in the AASM guidelines, which are not completely unambiguous [57]. Different versions of these guidelines exist, and different sleep centers may choose to abide by different sets of rules, or apply them slightly differently. The overall agreement between expert sleep scorers using the AASM guidelines is about 82.6%, with κ-score 0.76 [58, 59]. Hence, sleep staging performances above this level can be considered 'overfitting' to a certain sleep scorer or dataset.

1.4 Scientific knowledge gaps to address

Different deep learning architectures that adopt the sequence-to-sequence framework have shown sleep staging performances on par with expert sleep scorers on large PSG datasets with mainly healthy participants. These results may lead us to believe that sleep staging is a solved problem, but the absence of automated sleep staging algorithms in clinical practice suggests otherwise. Although automatic sleep staging has come a long way, there are still challenges to overcome [41].

Validity in clinical populations

First, most sleep staging methods have been developed and tested on public datasets consisting of healthy, young adults. Databases with **older populations and sleep pathologies** are rarer, and sleep staging performances are still lacking on such diseased populations [41]. Although sleep is more difficult to score in diseased populations, partly because of their more fragmented sleep, these populations are also most relevant for clinical application of the algorithms. Hence, it is essential to demonstrate the clinical utility of automated algorithms in such clinical populations. An accompanying challenge involves determining when sleep staging is 'good enough' for these groups, given that sleep disturbances and pathologies are reported to lower the inter-rater agreement [57, 59, 60].

Domain shift

Second, deep learning methods trained on one dataset may struggle to **generalize** to another dataset, due to the so-called domain shift problem. This mismatch in the statistical data distribution can have many causes, including a different institution, scorer, device, modality, disorder, acquisition condition, or demographic. It breaches the assumption that training and testing data are independent and identically distributed (i.i.d.), a requirement for machine learning systems to deliver their expected performance. A naive solution would be to compile training data from varied sources, covering all conditions that can be anticipated at deployment time [47, 56]. However, this solution would be costly and impractical. Additionally, training a new deep learning model from scratch for each new dataset is burdensome, as we will further discuss in the next paragraph.

Wearable devices

Third, closely linked to the domain mismatch issue, is the challenge of automated sleep scoring for wearable EEG devices. In fact, we can state that the domain mismatch issue is the most severe for wearable devices, as these typically acquire new, non-standard EEG modalities. This leads to a large channel mismatch with standard EEG or PSG data, such that deep learning models trained on such hospital data can not directly be applied to wearable EEG data. The domain mismatch issue was already discussed in the previous paragraph, but wearables have some unique characteristics that pose additional challenges for training deep learning models.

First, labeled datasets for mobile EEG setups are typically small, usually including only 10 to 20 participants. This limitation stems from the difficulty in obtaining reliable annotations for wearable devices. The differing electrode configurations and lower signal quality compared to PSG make these data challenging for sleep physicians to interpret [35]. The only way to obtain accurate expert labels for wearable data is to simultaneously record brain activity with PSG and wearable recording devices and use the manual labels on PSG as ground truth for the wearable data. The resulting scarcity of labeled data necessitates new techniques for processing data from wearable devices that can learn effectively from **smaller labeled datasets**.

On the other hand, unlabeled data can cheaply and easily be acquired with wearables. Large-scale sleep monitoring with wearables will lead to a rapid **increase in unlabeled data**. Additionally, longitudinal monitoring presents a valuable opportunity to enhance sleep staging accuracy through personalized

sleep scoring. In conclusion, deep learning strategies need to be adapted and optimized for wearable use cases, where they should exploit both small annotated datasets and larger unannotated datasets.

End-user trust

The fourth issue limiting the usability of deep learning algorithms for sleep staging is the lack of trust from clinicians, which can be largely attributed to deep learning's black-box nature. This issue highlights a broader limitation of deep learning that restricts its application across various medical fields, despite achieving human-expert-level accuracy in controlled tasks [46, 61–63]. As deep neural networks are highly complex, humans can not fully understand how they map inputs to outputs. Considering the impact on human lives in clinical decision-making, medical professionals are understandably hesitant in relying on such black-box systems. To integrate automated sleep scoring algorithms into clinical practice, their **transparency** needs to be improved.

The significant subjectivity and inter-rater disagreement among human sleep scorers present an extra hurdle in establishing trust in automated methods. As a model can be considered the twin to the human who annotated the training data, it will encounter disagreement in sleep staging predictions with the new clinician who is its end-user. To ensure acceptance by clinicians, the model must **adapt** to the scoring preferences of each new scorer. Once satisfactory sleep staging performance has been achieved by addressing the first three challenges, overcoming the trust issue represents the final step toward integrating automated sleep staging methods into clinical practice.

1.5 Research objectives and overview

1.5.1 Research objectives

The general research objective of this PhD thesis is to develop machine learning frameworks that facilitate the adoption of automated sleep scoring methods in clinical practice, with the future of wearable-based sleep monitoring in mind. More specifically, the main goal is to develop artificial intelligence (AI) methods on both PSG and wearable EEG data to reliably measure sleep in clinical cohorts, with a particular focus on patients with AD. The subobjectives of this thesis are threefold and aim to address the knowledge gaps identified in Section 1.4.

1. The first subobjective is to develop methods for improving the generalization of sleep stage classification models to new data, using transfer learning and domain adaptation approaches. The emphasis is on improving automated sleep staging for wearable devices, in which the domain shift problem is most problematic. However, these methods are equally applicable to sleep staging in PSG datasets with different characteristics.

   Addressed challenges: domain shift and wearable devices.

2. The second subobjective is to develop methods to improve the transparency and user interaction of sleep staging models by use of uncertainty estimation. These methods should be applied to hospital-based PSG scoring, where manual scoring remains standard practice due to transparency and trust issues. Additionally, I want to investigate how uncertainty estimation can help improve models for wearable remote sleep monitoring.

   Addressed challenges: end-user trust.

3. The third subobjective is to prove the clinical validity of automated sleep staging methods by deploying them on clinical populations of interest and exploring the clinical relevance of the derived sleep parameters. The main clinical population under study is AD.

   Addressed challenges: validity in clinical populations.

The first two subobjectives introduce technical innovations from a machine learning perspective, whereas the third subobjective is primarily focused on the clinical application. Ultimately, I envision for my algorithms to contribute toward the future of sleep medicine by enabling the integration of automated sleep staging into clinical practice. With a focus on wearable EEG devices, the primary aim of this thesis is to facilitate the study of sleep-wake disturbances on a population level.

1.5.2 Chapter overview

This thesis is subdivided into three main parts, each tackling one of the project's subobjectives. After the introductory Part I, Part II focuses on improving the generalization of automated sleep stage classification models, aiming to achieve robust models for wearable EEG measurements. After these performance issues are addressed, Part III focuses on methods to address the transparency issues through the use of uncertainty estimation. To address the challenge of clinical utility, all methods in Part II and III are developed and tested on clinical populations to demonstrate their usefulness in clinical practice. Part IV

presents a larger advancement regarding the clinical utility of automatic sleep staging methods. It demonstrates the application of sleep staging in various clinical cohorts and extends beyond traditional sleep staging to reveal the diagnostic potential of the derived sleep parameters. Finally, Part V describes the conclusions and directions for future research. The five parts and the corresponding chapters are illustrated in Figure 1.5.

Five of the six chapters in Part II-IV are each based on one journal paper. Each of these publications is first-authored by myself, under supervision of prof. M. De Vos. Close collaborators during my PhD have been dr. H. Phan (currently at Amazon, AGI) for Chapter 3-4, and N. Seedat and prof. M. van der Schaar (University of Cambridge) for Chapter 5-6. The latter collaboration was the result of a three-month research visit to the University of Cambridge. Chapter 8 is the result of a close collaboration with A. Devulder, and is hence co-first-authored by us both. Chapter 7 is not based on a first-author publication, but presents highlights of different clinical collaborations. For this chapter in particular and for the clinical datasets used throughout this thesis, I want to give special thanks to my clinical collaborators: P. Borzée, A. Devulder, L. Van den Bulcke, L. Micca, R. Proost, J. Macea Ortiz, prof. dr. Testelmans, prof. dr. B. Buyse, prof. dr. M. Van Den Bossche, prof. Gilat, and my co-supervisor prof. dr. W. Van Paesschen.

Part I - Introduction

Chapter 1 is the present chapter, introducing the important concepts, problem statement, and research objectives of this thesis.

Chapter 2 presents the materials that are used in the subsequent chapters, detailing the main datasets collected and used, and the neural network architectures employed.

Part II - Transfer learning approaches to improve generalization

Chapter 3 explores the development of a robust sleep staging methodology for sleep staging with a **limited set of labeled training data**. A transfer learning approach is used to compensate for the small size of the annotated datasets. The method is trained and validated using small, annotated datasets with simultaneous recordings of PSG and wearable EEG. To ensure this approach generalizes across many settings, the datasets include data from different wearables and demographics.

Chapter 4 describes a second way to improve sleep staging performances on wearable EEG data, by use of **large amounts of unlabeled data**. As wearable devices allow to conveniently record unlabeled data in a home environment, these can be exploited with unsupervised or semi-supervised machine learning methods. A semi-supervised transfer learning method is developed based on a method called adversarial domain adaptation [64]. It is again validated on multiple different wearables.

Part III - Uncertainty estimation to increase trust

Chapter 5 explores U-PASS, a novel framework which increases the transparency of machine learning methods for sleep staging by **providing estimates of uncertainty**. These uncertainty estimates can be beneficial to choose which data to use for training, they can be used to interact with the clinician through active learning approaches, and finally, they can help to visualize uncertain machine learning predictions for the user at deployment time. The framework is tested on a clinical PSG dataset.

Chapter 6 extends this novel framework to the use case of wearable devices, deploying **U-PASS in a wearable EEG sleep staging task**. Through a human-machine interaction experiment with manual review, this chapter investigates whether a clinician is able to correct the uncertain predictions. It also explores the clinical utility of deep learning based on wearable EEG, checking whether the algorithm picks up on relevant diagnostic features.

Part IV - Validation in clinical populations

Chapter 7 demonstrates the clinical validity of automated sleep staging methods by deploying them on different clinical populations of interest. It investigates if the sleep parameters derived with automated methods correlate with clinical parameters in those populations.

Chapter 8 aims to go beyond sleep staging, and investigates how automated pipelines can be useful for disease screening. In particular, this chapter explores if sleep recordings with a wearable device can be used to non-invasively screen for AD. An AI model for AD detection is developed based on sleep features extracted through automated sleep staging on wearable EEG recordings. Consequently, we identify the sleep parameters that are most informative for AD detection.

Part V - Conclusion

Chapter 9 recapitulates the main findings of this thesis in relation to the defined challenges and research objectives. Furthermore, it presents several avenues for future research and implications for the future of sleep medicine based on the findings from this thesis.

Appendices

Appendix A presents preliminary work exploring the intersection between uncertainty estimation and explainability, to further increase the transparency of deep learning models in clinical applications. Proof-of-concept experiments are performed to investigate if explainability methods could be harnessed to explain sources of uncertainty.

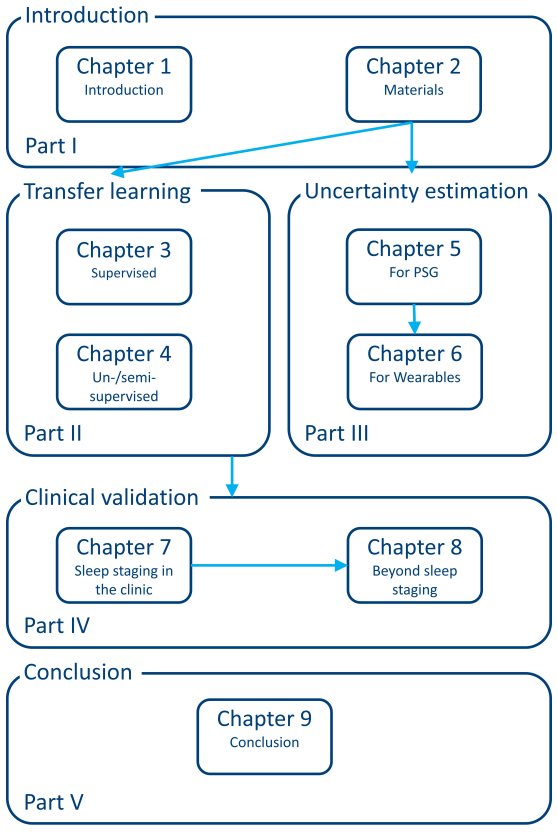


Figure 1.5: A schematic overview of the thesis, representing the different chapters and the links between them.

2 | Materials

ABSTRACT | This chapter provides an overview of the materials used in the subsequent chapters (Chapter 3-8). This includes a description of the main datasets collected and utilized throughout my thesis, as well as an outline of the neural network architectures and signal processing tools employed. The present chapter is thus largely based on the materials and methods sections from publications that make up the next chapters. Only datasets utilized in multiple chapters are described here, while datasets used in a single chapter are detailed in their respective chapters to enhance the readability of the thesis.

2.1 Datasets

Datasets with sleep recordings are essential to develop and validate our methods. With the research objectives outlined in Chapter 1 in mind, I opted to use a variety of datasets, including different wearable devices and participants of varying ages and health conditions. A summary is shown in Table 2.1.

2.1.1 Healthy

First, I used a number of public datasets consisting of mainly healthy participants: a large PSG dataset and two smaller labeled datasets collected with two different wearables.

Montreal Archive of Sleep Studies

The Montreal Archive of Sleep Studies (MASS) is a large, public PSG database [65]. It consists of 200 laboratory-based PSG recordings of 97 men and 103 women aged between 18 and 76 years old, recorded at three different hospital-based sleep laboratories of the Center for Advanced Research in Sleep Medicine (CARSM) in Montreal, Canada. All recordings are sampled at 256 Hz and feature 4 to 20 EEG channels, EOG, ECG, EMG, and respiratory signals.

Sleep stages were scored according to either the R&K guidelines [66] or the AASM standard [67]. As the R&K guidelines prescribe 20-second segments and six sleep stages, we align the R&K portions of the dataset with the AASM guidelines as described in [46]: we combine the scorings into the five sleep stages of the AASM standard {W, N1, N2, N3, and REM} and convert all segments into 30-second ones. 20-second segments are expanded by padding them with slices of 5 seconds from both neighboring segments.

Surrey – cEEGrid

The Surrey - cEEGrid database [35,68] was recorded at the University of Surrey from 12 healthy adult volunteers, who were mostly good sleepers. Full-night wearable EEG recordings and PSG recordings were simultaneously collected. Wearable EEG measurements were obtained with the cEEGrid array, a wearable

EEG recording device consisting of a flexible printed electrode strip around the ear [30, 69]. One cEEGrid electrode strip, comprising 10 electrodes, was positioned behind each ear. The cEEGrid data were recorded with a wireless SMARTING amplifier (mBrainTrain, Belgrade, Serbia) and a Sony Z1 Android smartphone at a sampling rate of 250 Hz. Manual annotation was based on the PSG [35].

Dreem – Headband

The Dreem – Headband dataset [36] was recorded at the French Armed Forces Biomedical Research Institute's (IRBA) Fatigue and Vigilance Unit (Bretigny-Sur-Orge, France). It is comprised of recordings from 25 adult volunteers with self-reported quality of sleep varying between no complaints and sub-threshold insomnia symptoms [70]. EEG signals were recorded with the Dreem Headband, a reduced-montage wearable dry-EEG device recording from five frontal and occipital EEG electrodes. This dataset also includes simultaneous PSG measurements, which are scored with a consensus based on the manual labels of five sleep experts [36].

2.1.2 Diseased

Secondly, we collected two new datasets from relevant clinical populations: older participants with OSA and dementia. To investigate the potential for home-based monitoring, data were collected with both PSG and a wearable device.

Leuven – SensorDot Senior

The Leuven – SensorDot Senior (SDS) sleep database consists of sleep recordings from patients of the sleep laboratory at UZ Leuven, which we collected from 2019 to 2022. Initially, 100 patients were included, but 10 recordings were lost due to missing or faulty data. The final dataset comprises 90 participants between 60 and 79 years old with suspected sleep apnea.

The dataset encompasses measurements of overnight PSG, simultaneously acquired with measurements from a wearable device. The wearable device is a two-channel behind-the-ear EEG recording device (Figure 2.1), called the SensorDot (Byteflies, Antwerp, Belgium). It measures EEG, ACM, and angular velocity, at sampling rates of 250 Hz, 50 Hz, and 50 Hz, respectively. It consists of a small (24.5 x 33.5 x 7.73 mm; 6.3 g) biopotential amplifier. It is placed

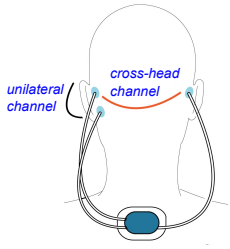


Figure 2.1: Main setup for the EEG measuring wearable device SensorDot in this thesis.

in the neck, with one to two electrodes behind each ear, creating unilateral or cross-head EEG channels. These wearable EEG derivations have previously successfully been employed for focal epileptic seizure detection [71, 72]. In the first 48 recordings, two unilateral channels were recorded. However, we found that the unilateral EEG channels were unsuitable for sleep staging. As a result, we recorded the next 52 recordings with one unilateral and one cross-head channel (see Figure 2.1). With the lost data, this resulted in a dataset consisting of 45 recordings with two unilateral wearable EEG channels and 45 recordings with a unilateral and cross-head EEG channel.

PSG recordings were acquired with Medatec software (Medical Data Technology, Braine-le-Château, Belgium), using a sampling rate of 200 Hz, and following the recommended AASM guidelines version 2.6 [7]. The PSG recordings were manually scored conforming to the same guidelines. With the hospital equipment, we created two additional EEG channels to mimic the wearable EEG recordings by adding an extra EEG electrode behind the right and left ears, both referenced to A1 (located at the left ear). These channels were used to synchronize the PSG and wearable recordings using cross-correlation as a similarity metric. In the 45 recordings with two unilateral wearable EEG channels, the cross-head EEG acquired in this manner substituted for the wearable EEG in our experiments.

This dataset was acquired in accordance with the Declaration of Helsinki and approved by the Ethical Committee of University Hospitals Leuven. Informed consent was obtained from all participants or their caregiver. The clinical trial number was NCT04755504 (S64190/B3222020000148).

Leuven – SensorDot AD

The Leuven – SensorDot AD (SDAD) dataset comprised patients with AD and control participants from a cross-sectional study conducted at UZ Leuven, with recruitment spanning from 2019 to 2022 [73]. The participants, aged between 55 and 85 years, were divided into a group of 30 control subjects and 35 patients with either probable clinical AD (n = 7) or biomarker proven AD dementia (n = 28) [74,75]. Details are shown in Table 8.B. Thirty-four percent of patients with AD had an MMSE $\geq$ 27/30 [76] and were further classified as being in the prodromal AD stage. Control subjects in this dataset were defined as healthy volunteers, free of subjective cognitive decline, with an MMSE $\geq$ 27/30, and a Clinical Dementia Rating scale (CDR) score of 0 [77].

Like the SDS dataset, the SDAD dataset consists of simultaneous measurements of PSG and recordings with the SensorDot wearable device. The SensorDot recorded one unilateral and one cross-head EEG channel. PSG recordings were conducted at-home with a slightly different setup from the AASM guidelines: scalp electrodes were placed using the International 10-20 system with additional lower temporal electrodes (F9, F10, T9, T10, P9, P10) without P3 and P4, and Cz was used as the ground. Additional sensors for PSG included EOG1/2, chin EMG, nasal flow (NAF2P), thoracic and abdominal resistance bands, finger pulse oximetry, ECG derivation Eindhoven I and EMG of one leg. Finally, a camera with audio recorder was installed in the patient's bedroom for nighttime activity recording. In this dataset, electrodes P9/10 were used as reference electrode instead of A1/2, and apneas were scored based on the flow signal on NAF2P. The PSG recordings were scored from lights-off until lights-on conform the AASM scoring rules version 2.6 [7]. The presence of sleep apnea was defined as an apnea-hypopnea index (AHI) $\geq$ 15/h.

This study was conducted in accordance with the Declaration of Helsinki and approved by the Ethical Committee of University Hospitals Leuven. Informed consent was obtained from all participants or their caregiver. The clinical trial number was NCT03617497 (S61745).

Table 2.1: Summary of the datasets used in this thesis.

Dataset	Nb. recordings	Subject characteristics	PSG	Wearable	Ref.	Chapters
MASS	200	Healthy, age 18-76	In-laboratory, varying equipment	/	[65]	3,4,7
cEEGrid	12	Healthy, age 18-63	In-laboratory, SomnoHD	Around-the-ear	[68]	3,4
Dreem	25	Healthy, age 23-50	In-laboratory, Siesta 802	Headband	[36]	4,7
SDS	90	Suspected OSA, age 60-79	In-laboratory, Medatec	Behind-the-ear		3,4,5,6,8
SDAD	35 30	AD, age 56-83 Older controls, age 55-80	At-home, Medatec (ambulatory)	Behind-the-ear	[73]	7,8

2.2 Sleep staging networks

After having introduced deep learning methods for sleep staging in Section 1.3.3, this section provides an overview of the network architectures used in this thesis. Formally, sleep staging consists of classifying a sequence of PSG or EEG epochs $(x_1, x_2, ..., x_M)$ into the corresponding sequence of outputs $(y_1, y_2, ..., y_M)$ by maximizing the conditional probability $p(y_1, y_2, ..., y_M|x_1, x_2, ..., x_M)$. All sleep staging experiments in this thesis are performed with one of two different neural network architectures: a compact three-layer attention-based recurrent neural network (ARNN) [46] on the one hand, and a state-of-the-art sleep staging network, SeqSleepNet [46] on the other hand. Both are illustrated in Figure 2.2. The networks are trained to classify 30-second segments of recorded data into the five sleep stages according to the AASM guidelines [67]. Both networks are capable of handling varying numbers of input channels. Before being presented to these networks, all the recorded data undergo identical pre-processing procedures.

2.2.1 Pre-processing

The recorded signals are pre-processed to facilitate their analysis. First, they are filtered to remove any unwanted noise outside the desired frequency ranges and resampled to a sampling rate of 100 Hz. Filtering is performed with a finite impulse response (FIR) bandpass filter, only letting through frequencies from 0.3 to 40 Hz for the EEG, EMG, and EOG signals, and from 0 to 12.5 Hz for the ACM and angular velocity signals.

Next, the signals are transformed to visualize how their frequency content changes over time. This is done using a method called the short-time Fourier transform (STFT), which analyzes the signals in short segments of 2 seconds each, with a 1-second overlap between segments, and uses a specific type of windowing function called a Hamming window to minimize signal distortion. Additionally, this transformation includes logarithmic scaling, which compresses the wide range of values into a more manageable scale. This transformation results in a visual representation called a spectrogram, which shows how the frequency content of the signals varies over time. The resulting spectrogram is normalized so that its values have an average of zero and a standard deviation of one.

2.2.2 Attention-based recurrent neural network

The simple ARNN model (Fig. 2.2(a)) [46] classifies sleep stages with a one-to-one classification scheme, meaning it takes a single epoch (x) and outputs the corresponding sleep stage (y). It consists of three layers:

- Filterbank layer: This layer filters the frequency dimension using weights that the model learns during training.
- Bidirectional RNN (biRNN): This layer allows for sequential modeling and processes the temporal information within a 30-second segment. It is implemented with a Gated Recurrent Unit (GRU) cell [78], a gating mechanism that allows to forget or retain certain information.
- Attention layer: This layer combines the vectors extracted by the biRNN into a single feature vector, which represents the segment.

The final classification is done by passing this feature vector through a fully connected layer with a softmax activation function. The softmax activation function makes sure the output is mapped to a probability distribution between 0 and 1. Hence, for every input segment, the network returns a probability for every class (W, N1, N2, N3, R). The entire network is trained by minimizing the cross-entropy loss, which measures the difference between the predicted and actual sleep stages.

2.2.3 SeqSleepNet

SeqSleepNet (Fig. 2.2(b)) [46] follows a many-to-many classification scheme, meaning it takes multiple segments as input and predicts the corresponding sleep stages for all of them simultaneously. As such, it transforms a sequence of M segments $(x_1, x_2, ..., x_M)$ into the corresponding sequence of M sleep stages $(y_1, y_2, ..., y_M)$. Throughout this thesis, we use a sequence length $M = 10$ (the equivalent of 5 minutes). In the original SeqSleepNet paper [46], larger sequence lengths such as $M = 20$ and $M = 30$ were also explored, but they resulted in only marginal improvements while significantly slowing down the training process.

SeqSleepNet uses the ARNN architecture as its first block to generate a feature vector for each segment. These feature vectors are then passed to a second biRNN layer, which models the relationships between the segments in the sequence. This sequence-level biRNN layer works similarly to the segment-level biRNN layer, processing the entire sequence of M feature vectors and producing M output vectors, one for each segment.

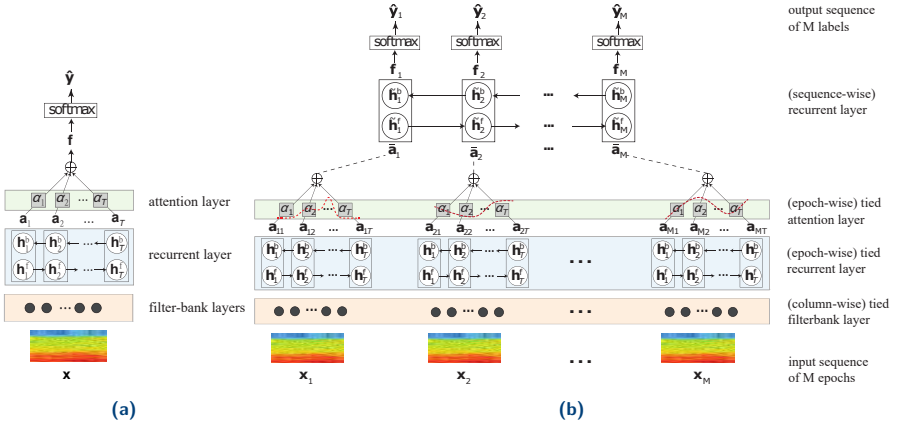


Figure 2.2: Illustration of the two sleep staging networks used throughout this thesis. (a) The ARNN, performing single segment-based sleep staging, (b) SeqSleepNet, designed for sequence-to-sequence classification. Both figures are adapted from [46].

Finally, these M output vectors are classified into M sleep stages by a fully connected layer with a softmax activation. The network is trained in an end-to-end manner, by minimizing the average cross-entropy loss over all segments in the sequence.

To train SeqSleepNet on all the possible sequences in a dataset, we sample sequences from the dataset with a shift of one segment, i.e. an overlap of $M-1$ segments. At test time, sampling the test set with the same shift results in an ensemble of M predictions for every segment. These predictions are aggregated by summing the logarithmic posterior probabilities over the ensemble. For further details of both networks, please refer to [46].

2.2.4 Parameters and settings

The ARNN and SeqSleepNet are both implemented with the *Tensorflow* framework [79]. Training parameters and optimization strategies follow the specifications from the original paper [46], including the Adam optimizer with a learning rate of 10^{-4} and L2 regularization.

2.2.5 Rationale behind the architecture choice

As will be explored in the following chapters, all the algorithms and methods developed in this thesis are model-agnostic, meaning they should work with any type of sleep staging model. Among the numerous possible sleep staging architectures discussed in Section 1.3.3, SeqSleepNet was primarily used for two main reasons.

First, maintaining consistency in the network architectures allows for more straightforward performance comparisons across different datasets and for the developed methods. Second, these two architectures have a relatively small number of parameters. SeqSleepNet, which is based on two RNN layers, has only around 10^5 parameters. In contrast, other state-of-the-art architectures, often based on CNNs, have significantly higher parameter counts, making them more prone to overfitting and computationally demanding to train. In a recent study[1], we analyzed the performance and parameter count of several state-of-the-art network architectures. SeqSleepNet emerged as both one of the smallest and best-performing networks.

The ARNN is just the epoch encoder of SeqSleepNet. Since it does not benefit from sequence-to-sequence modeling, it achieves lower sleep staging accuracies. It was used primarily as a baseline to determine whether both complex sequence-to-sequence and simple epoch-to-epoch classification architectures could benefit from some of the developed approaches.

[1] H. Phan, E. R. M. Heremans, O. Y. Chén, P. Koch, A. Mertins and M. De Vos, "Improving Automatic Sleep Staging Via Temporal Smoothness Regularization," in Proceedings of the *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, pp. 1-5, 2023.

Part II

Transfer learning approaches to improve generalization

3 | Feature matching as an improved transfer learning technique

This chapter is largely based on **E. R. M. Heremans**, H. Phan, A. H. Ansari, P. Borzée, B. Buyse, D. Testelmans, and M. De Vos, "Feature matching as improved transfer learning technique for wearable EEG," *Biomedical Signal Processing and Control*, vol. 78, pp. 104009, 2022.

ABSTRACT | Wearable sleep monitoring devices require automated sleep staging algorithms that can learn from small amounts of labeled data. Transfer learning has the ability to adapt neural network weights from a source modality (e.g. standard electrode configuration) to a new target modality (e.g. non-conventional electrode configuration). We propose feature matching, a new transfer learning strategy, as an alternative to the commonly used fine-tuning approach. This method uses limited amounts of paired samples of the source and target modality, extracting features of the target modality and matching these to the features from the corresponding samples of the source modality. We compare feature matching to fine-tuning for three different target domains, with two different neural network architectures, and with varying amounts of training data. Particularly on small cohorts (i.e. 2 - 5 labeled recordings), feature matching systematically outperforms fine-tuning with mean percentage point improvements in accuracy ranging from 0.3% to 3.0% for the different scenarios and datasets. Our findings suggest that feature matching outperforms fine-tuning as a transfer learning approach, especially in very low data regimes. As such, we conclude that feature matching is a promising new method for wearable sleep staging with novel devices.

3.1 Introduction

The size of wearable EEG datasets is often very small, which greatly limits the performance of supervised learning methods. To compensate for the lack of data, automated sleep staging algorithms for wearable EEG could exploit information extracted from large, manually labeled datasets with standard EEG modalities. After pre-training a model on a large dataset of a standard EEG modality, transfer learning can be used to transfer the learned information to improve the sleep staging performance on a small dataset recorded with a new modality. Previous studies [47, 80, 81] already used this principle for sleep staging applications, and showed that transfer learning successfully deals with the channel mismatch between different EEG channels (and even across modalities between EEG and EOG). These earlier investigations made use of the simple fine-tuning approach, in which sleep staging networks were trained on a large labeled dataset of a source modality, and then fine-tuned on a small dataset of a target modality. In [82], fine-tuning was used with Kullback-Leibler (KL) divergence regularization for personalization to specific subjects.

Although the fine-tuning approach is useful, it has some limitations. First, the model forgets the source domain when it is fine-tuned on the target domain. Therefore, this method does not really allow to study the relationship between the data representations of these different domains. Second, traditional fine-tuning approaches use purely supervised learning. Therefore, these methods cannot easily be extended to include unlabeled data in the fine-tuning step itself, motivating the need to explore alternative methods.

In the field of deep domain adaptation, strategies like domain mapping and domain-invariant feature learning have been developed to align features between source and target domains [83]. Various supervised domain adaptation techniques for producing domain-invariant features have been successful in achieving better generalization across datasets in image processing applications [84–87]. However, to our best knowledge, such supervised domain adaptation techniques have not been employed in EEG applications yet.

Additionally, in the field of multi-view learning, common representation learning methods like canonical correlation analysis (CCA) [88], correlational neural networks (CorrNet), and bridge neural networks (BNN) have been used to achieve correlated representations of two different modalities [89–93]. These common representation learning methods were built with several downstream tasks and applications in mind, such as transfer learning, matching of

corresponding samples across different views, and reconstruction of missing views (only CorrNet and BNN). In CCA and CorrNet, the correlation between feature representations is minimized [89–92]. BNNs [93] use Euclidian distance between feature representations, and construct both negative and positive pairs of samples. All these methods focus on aligning two feature spaces and are primarily evaluated by checking correlation measures between the two views. The learned feature embeddings are then sometimes used in a second step to perform downstream tasks like classification, but the encoders are not simultaneously optimized to perform both feature alignment and classification.

In this chapter, we propose a novel, improved transfer learning technique (see Figure 3.1) for automated sleep staging. Our method aligns the feature space of a new EEG modality (the target domain) with that of a standard EEG modality (the source domain). Importantly, we make use of a separate encoding network for the source modality and target modality, a choice which is movitated by the large domain mismatch between recordings from different electrode positions. Separate encoders allow for more flexibility to learn features separately for both domains. Similar to domain adaptation techniques in image processing [84–87], our method aligns the source and target domains whilst simultaneously training a classifier for the prediction task. However, unlike these approaches and the fine-tuning method, our method not only leverages label information but also uses the correspondence between paired data samples across both modalities. It hence optimizes a paired matching loss like the common representation learning methods, while simultaneously training the encoders for classification. Labeled datasets with wearable EEG inevitably have simultaneous recordings of the standard EEG (source modality) and wearable EEG (target modality), because a standard EEG recording is needed to obtain ground truth labels for wearable EEG [35]. Therefore, the use of paired samples from simultaneous recordings is not a limiting factor.

Figure 3.1 illustrates some of the main concepts of this chapter that were touched upon in this introduction. The remainder of this chapter is structured as follows. Section 3.2 describes the datasets used to develop our method and to train and test the classifiers, as well as the neural network architectures that were used for sleep staging. Section 3.3 gives a short introduction on transfer learning, and explains the proposed feature matching method. Then, Section 3.4 reports on the experiments conducted to validate our method and to compare it to the state-of-the-art fine-tuning approach, and Section 3.5 shows the obtained results. Finally, Section 3.6 discusses the benefits of transfer learning and the advantages of our method compared to fine-tuning, and Section 3.7 presents the main conclusions.

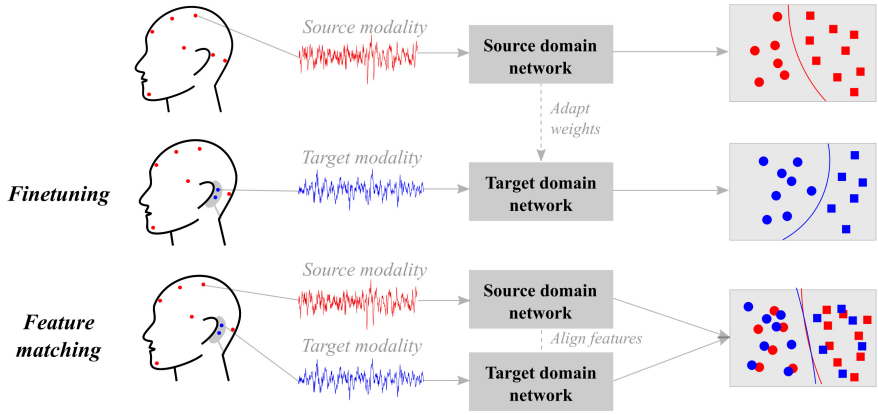


Figure 3.1: The conceptual difference between fine-tuning and feature matching. Top: (pre-) training of a neural network to perform a classification task on the source domain. Middle: in fine-tuning, the network weights are adapted by consequently training on the target domain. Bottom: in feature matching, the weights are adapted by training on both the source and the target domain and matching the features of both domains with each other.

3.2 Materials

3.2.1 Source domain

MASS – C4 The MASS database [65] (see Section 2.1.1) is selected as the source domain, as it is a large, public PSG dataset. The standard C4-A1 EEG channel from this database is selected as the source domain. This EEG derivation is a good choice for the source domain, as it is commonly used in automated sleep staging methods [46, 48, 94–97] and recommended in sleep scoring guidelines [67, 98].

3.2.2 Target domains

For the target domain, we use three different modalities from distinct datasets to test our method extensively in different scenarios.

MASS – EOG The first target domain is the mean EOG (EOG left-right) of the MASS database. This target domain allows us to investigate transfer learning from one modality to another one, within one database and population. Forehead electrodes have shown promise for wearable sleep monitoring systems [34,99,100]. Our approach can thus be used for sleep staging on a single EOG channel, which is beneficial for wearable monitoring.

Surrey – cEEGrid The Surrey – cEEGrid database [35,68] (see Section 2.1.1) serves as the second target domain. From this dataset, we use the right-ear front-versus-back derivation (FB(R)). This second target domain allows to investigate transfer learning to a dataset of real wearable data, acquired in a completely different sleep laboratory.

Leuven – SensorDot Senior The third target domain consists of the first 28 sleep recordings of the Leuven - SDS database (see Section 2.1.2). We use the cross-head behind-the-ear EEG channel of this dataset as a simulation for a wearable EEG channel. The elderly and diseased population in this last target domain introduces an extra source of domain mismatch. It reflects a realistic use case for ambulatory sleep monitoring and allows to investigate transfer learning to a different dataset and modality.

3.2.3 Sleep staging networks

The experiments are performed with the simple ARNN and SeqSleepNet [46], both introduced in Section 2.2. Both networks can cope with different numbers of input channels, but in this chapter, we focus on single-channel sleep staging.

3.3 Feature matching

3.3.1 Transfer learning framework

Transfer learning is defined as the transfer of information between two different but related machine learning problems [101]. It is aimed at addressing machine learning scenarios where a model is trained to perform a task on a source domain, but then applied to a potentially different task in a potentially different target domain. Formally, a domain consists of a feature space $\mathcal{X}$ and a marginal probability distribution $P(X)$: $\mathcal{D} = \{\mathcal{X}, P(X)\}$. A task consists of a label space $\mathcal{Y}$ and predictive function $f(.)$ projecting the input space onto the label

space: $\mathcal{T} = \{\mathcal{Y}, f(.)\}$. When a mismatch between two tasks or domains occurs ($\mathcal{T}_S \neq \mathcal{T}_T$ or $\mathcal{D}_S \neq \mathcal{D}_T$ with subscripts S and T denoting source and target), transfer learning is used to overcome this disparity, improving $f_T(.)$ using information from $\mathcal{T}_S$ and $\mathcal{D}_S$ [83, 101, 102].

Applying these concepts to sleep stage classification on wearable EEG recordings, the source domain consists of a standard EEG channel of a large public sleep database. We aim to transfer knowledge from this domain to the target domain consisting of a small database with a non-traditional EEG channel. The task in both domains is sleep stage classification into the five sleep stages.

3.3.2 Feature matching method

Similar to the fine-tuning approach, we first pre-train a sleep staging network on a large sleep database, i.e. the source domain. Then, we use our novel feature matching method as a transfer learning approach to adapt this network to the small database with a new modality, i.e. the target domain. Whereas the fine-tuning approach only uses the labels of the target modality, our feature matching approach also exploits the correspondence between the simultaneously recorded data of both the source modality and target modality. We explicitly match the extracted feature vectors of the target modality with those of the source modality for the corresponding samples.

The reasoning behind this approach, is that through pre-training on a large source dataset, we learn features from the source domain which are superior to those we can learn from a very small target dataset. By minimizing the distance between the features of the source modality and target modality, we use information from the source domain to improve the target domain features. Feature vectors represent a precise location in the feature space, whereas a label only designates a general area in the feature space.

Each sleep staging network is conceptually split into two components: a feature extractor consisting of all layers but the last one, and the last layer itself which is the classification layer. Both of these are trained end-to-end as one network, but perform different tasks. The feature matching method, as illustrated in Figure 3.2, consists of two steps:

1. Initialization: make two duplicates of a sleep staging network architecture (see Figure 3.2), and initialize them both with the network weights pre-trained on the source domain,

2. Adaptation of the two parallel networks, with each network trained on a different modality. The first network, the source network (consisting

of feature extractor S and classification layer S in Figure 3.2), is further trained on the source modality. The second network, the target network (feature extractor T and classification layer T), is fine-tuned on the target modality.

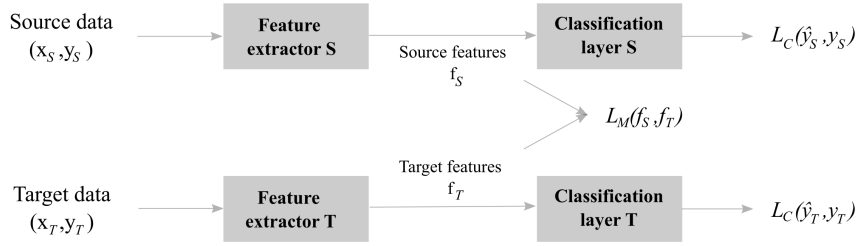


Figure 3.2: Feature matching consists of concurrently training two networks, while matching the feature representations.

Both the source and target modality networks get separately trained to classify sleep stages (network predictions $\hat{y}$) with data (x) and labels (y) from their respective modality. At the same time, their feature extractors are trained to minimize the mean squared error (MSE) between extracted features (f) of corresponding samples of the source modality and target modality. This training process is what we refer to as feature matching. The combination of both the source modality network and target modality network thus gets trained with the following loss function:

$$\mathcal{L} = \mathcal{L}_C(\hat{y}_S, y_S) + \mathcal{L}_C(\hat{y}_T, y_T) + \lambda_1 \mathcal{L}_M(f_S, f_T) + \lambda_2 L_2 \tag{3.1}$$

in which the subscript S and T designate the source modality and target modality, respectively. $\mathcal{L}_C$ is the cross-entropy or classification loss. $\mathcal{L}_M$ is the matching loss between the features of paired samples of the source modality and the target modality, computed with the MSE. L_2 is the L_2-norm of the network weights, used for regularization. The hyperparameter λ_1 determines the weight of the matching loss relative to the classification losses, and λ_2 determines the weight of the L_2 regularization term.

Alternatively to using the MSE loss between paired samples, the maximum mean discrepancy (MMD) loss could also be used to match the source and target features [103, 104]. However, this measure acts on a distribution level instead of using paired samples, not taking full advantage of the available knowledge in the case of simultaneous EEG recordings. In our experiments, the accuracy obtained with the MSE loss systematically outperformed the accuracy obtained when using the MMD loss.

The source code for the feature matching method is available at https://github.com/elisabethRMH/featurematching.

3.4 Experiments

3.4.1 Transfer learning scenarios

In order to validate our feature matching method and compare it to the state-of-the-art fine-tuning approach, we examined different transfer learning scenarios. In each scenario, the model was pre-trained on the C4-A1 derivation of a large dataset, and then adapted to overcome the channel mismatch with different target modalities from small datasets, each containing both the C4-A1 source derivation and the target modality.

We applied transfer learning to three different target modalities: EOG, cEEGrid, and cross-head behind-the-ear, and with two different network architectures: the compact ARNN network and SeqSleepNet. This resulted in a total of 6 transfer learning scenarios. In every scenario, transfer learning was performed with the classical fine-tuning approach, and with the novel feature matching method. In addition to the basic fine-tuning approach, fine-tuning with KL-divergence regularization as introduced in [82] was also added as an extra baseline for comparison. We also compared the transfer learning performances with simple training from scratch on the target domain, and with directly evaluating the networks trained on the source domain (we call this 'direct transfer').

3.4.2 Experimental setup

For every learning scenario, we trained the models on different amounts of target modality data to investigate how both transfer learning methods perform on smaller and larger datasets. Every experiment was performed as a modified k-fold cross-validation on the target dataset to obtain average performance measures. The modification with respect to normal cross-validation was made to control the size of the training set in every experiment. The modified training sets were subsets of each full training set. The sizes of these training subsets for transfer learning were of 10, 5 and 2 recordings. Table 3.1 shows the amount of recordings per dataset, the subdivision into a training set, test set and validation set for every round of cross-validation, and the number of recordings in the training subsets. Note that depending on the size of the dataset, there are multiple different possible subsets within one round's training set. For every round of cross-validation, we used 1 training dataset of 10 recordings, 2 training

sets of 5 recordings and 5 training sets of 2 recordings. The performance was thus evaluated on each fold for each training subset, and performance values were averaged over N evaluations (with N = number of folds ∗ number of subsets). Since the variability in performance is higher when smaller training sets are used, it makes sense to split the data into more different training sets when the training sets are smaller, to get a representative performance value.

Table 3.1: The datasets and their subdivision into training, validation and test sets for every cross-validation round. 'Training set' designates the total training set for a round, and 'training subsets' designates the modified training sets used for transfer learning.

Dataset	Channel	Number of recordings					Nb. folds
		Total	Training set	Training subsets	Val. set	Test set	
MASS - C4	C4-A1	200	180		10	10	20
MASS - EOG	EOG	200	180	10/5/2	10	10	20
Surrey - cEEGrid	cEEGrid	12	10	10/5/2	1	1	12
Leuven - SDS	Right ear-A1	28	24	10/5/2	2	2	14

The first set of experiments was carried out with the EOG from the MASS dataset as the target domain. In this case, the dataset of the source domain and target domain are actually the same. Pre-training was performed with 20-fold cross-validation on the C4-A1 recordings of the MASS dataset, with 180 recordings as a training set in every round. Then, transfer learning was performed with the same 20-fold cross-validation scheme, this time only using 10 EOG and C4-A1 recordings of the training set in each round of the cross-validation to simulate smaller datasets. Those 10 recordings were further subdivided into the aforementioned subsets for transfer learning (one subset of 10 recordings, two subsets of 5 recordings and five subsets of 2 recordings).

The experiments on the Surrey-cEEGrid dataset and the Leuven-SDS dataset follow a simpler scheme. In both of these cases, we pre-trained the networks on

Table 3.2: The distribution of the different classes in the three datasets. The total amount of 30-second segments in each dataset is shown, as well as the relative amount of samples for each sleep stage (in percentages).

Dataset	Total	Percentage of samples per sleep stage				
	nb. samples	W	N1	N2	N3	R
MASS	228870	13.56%	8.46%	47.15%	13.27%	17.55%
Surrey - cEEGrid	14598	38.34%	5.68%	31.60%	11.78%	12.60%
Leuven - SDS	30220	32.50%	9.35%	36.49%	10.56%	11.11%

all 200 C4-A1 recordings of the MASS dataset, except 10 recordings used as a validation set. Then, we performed the modified cross-validation on the two respective target datasets, using training subsets instead of the full training sets. For further details of the cross-validation procedure and subsets, we refer to Table 3.1. The sleep stage distribution in every dataset is shown in Table 3.2.

3.4.3 Minibatch construction

For feature matching, the minibatches are constructed in the following manner, illustrated in Figure 3.3. For the source modality network to retain its sleep staging capabilities on the source modality, it is presented with labeled source modality data from the source dataset (the MASS dataset in this case). In order to compute the feature matching loss, we need paired samples that have both modalities, i.e. the samples of the target dataset. The source modality network computes the features of the source modality data of the target dataset needed for this feature matching loss. The target modality network computes features of the target modality samples of the target dataset. Every minibatch thus consists partly of data from the source modality (used for $\mathcal{L}_C(\hat{y}_S, y_S)$ in the loss function 3.1), and partly of target modality samples with their corresponding source modality samples (used for $\mathcal{L}_M(f_S, f_T)$ and $\mathcal{L}_C(\hat{y}_T, y_T)$ in 3.1). In comparison, it should be noted that for simple fine-tuning, the minibatches are constructed using only target modality samples of the target dataset, as in this case, the model only trains with the classification loss on the target modality samples.

3.4.4 Training parameters

Networks are always pre-trained for 10 epochs and transfer learning (feature matching or fine-tuning) is performed for 20 epochs. During training, networks are evaluated on the validation set after every 200 training steps, and the best-performing network on the validation set is retained for evaluation on the test set. This acts as a regularization approach like early stopping. For pre-training and fine-tuning, the minibatches consist of 32 sequences. For feature matching, the size of minibatches (N_{mb}) is not constant across all training scenarios. The minibatches contain both 8 sequences from the target dataset ($N_{td,mb} = 8$) and a variable, larger number of sequences from the source dataset ($N_{sd,mb}$). The number of sequences from the source dataset in a minibatch depends on the proportion in size of the two datasets (N_{sd}/N_{td}), so that all the training samples of both datasets pass through the network once in every epoch: $N_{mb} = N_{td,mb} + N_{sd,mb} = 8 + 8 * N_{sd}/N_{td}$. This number is selected such that the minibatches fit within our GPU memory limit (with the GPU model NVIDIA

TU104 [GeForce RTX 2080]). In every training step, the classification losses and matching loss are computed by summing over the minibatch. As the matching loss L_M is summed over only 8 sequences of the minibatch, the matching loss weight λ_1 is fixed to $N_{mb}/8$ to give it the same relative importance as the source classification loss.

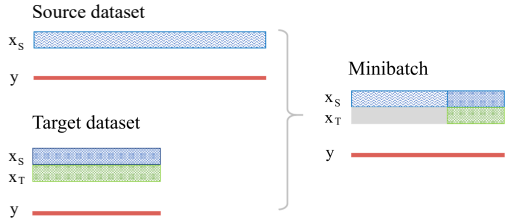


Figure 3.3: Illustration of a minibatch for training the feature matching approach. x_S indicates data of the source modality, and x_T indicates data of the target modality. y indicates the labels.

3.5 Results

3.5.1 Baseline performance on the source domain

First, to put our results in perspective, we show the network performances on the source domain. Training on the best possible EEG channel of a large dataset acquired from healthy subjects with standard hospital equipment should give an upper limit to what the networks can achieve with single-channel data. 20-fold cross-validation is performed on the source domain, with 180 recordings in the training set, 10 recordings as validation set and 10 recordings as test set for every cross-validation round. Table 3.3 shows the sleep staging performance of SeqSleepNet and ARNN on this dataset. It reports the accuracy, κ and wF1, as mean $\pm$ standard error over the 20 folds. The accuracy obtained with SeqSleepNet on single-channel data is 83.9%, which is in line with the results for multiple channels in [46].

3.5.2 Fine-tuning and feature matching performance on the target domains

We assess the performance of both transfer learning methods and compare it with direct transfer and training from scratch on the target domain. Table 3.3

shows the results for the three target domains: the EOG of the MASS database, the cEEGrid channel of the Surrey dataset, and the cross-head behind-the-ear channel of the Leuven - SDS dataset. The mean ± standard error is computed over all the cross-validation folds and training subsets as defined in Section 4.4.1. Figure 3.4 highlights some of the most important results, visualizing the difference in performance of the different methods for the three target domains.

The baseline performance on the large source domain is clearly higher compared to the performances obtained from training on smaller sized datasets of the target domains in Table 3.3. For every target domain, direct transfer performs worse than any other method tested. Training from scratch on the complete dataset performs worse than transfer learning on subsets when the dataset is small (e.g. the Surrey - cEEGrid dataset), but it performs better when the dataset is large (e.g. the MASS - EOG dataset). When comparing the two transfer learning techniques, feature matching always outperforms fine-tuning when 2 recordings of the target modality are used. When 5 recordings are used, feature matching also outperforms fine-tuning in most cases. When using 10 recordings, the difference is very small, so we could state that both approaches obtain a similar performance.

Table 3.3: Sleep staging performance of ARNN and SeqSleepNet on source and target domains, using feature matching, fine-tuning, fine-tuning with KL-divergence regularization, direct transfer and training from scratch. Mean ± standard error is set out for the accuracy (acc.), κ and wF1. # is the number of recordings in the training set for every scenario. The performance metrics are averaged over N values as described in the manuscript. For each transfer learning scenario, the best performance is in bold.

#	Method	(N)	ARNN			SeqSleepNet		
			Acc	κ	wF1	Acc	κ	wF1
Source domain: MASS - C4 dataset								
180	Scratch	(20)	80.1±0.5	0.717±0.007	79.6±0.4	83.9±0.4	0.769±0.006	83.6±0.4
Target domain 1: MASS - EOG dataset								
	Feature matching		**73.3±0.3**	**0.617±0.004**	**72.8±0.3**	**77.6±0.3**	**0.680±0.004**	**77.1±0.3**
2	Fine-tuning	(100)	71.9±0.4	0.603±0.005	71.7±0.4	76.1±0.4	0.656±0.005	75.5±0.4
	Fine-tuning with KL		73.0±0.3	0.614±0.004	72.5±0.3	77.5±0.3	0.671±0.004	76.4±0.3
	Feature matching		**74.9±0.4**	**0.642±0.005**	**74.5±0.4**	**79.7±0.3**	**0.709±0.005**	**79.2±0.4**
5	Fine-tuning	(40)	74.6±0.4	0.640±0.005	**74.5±0.4**	78.8±0.4	0.697±0.006	78.6±0.4
	Fine-tuning with KL		74.7±0.4	0.639±0.005	74.2±0.3	79.2±0.3	0.697±0.005	78.4±0.4
	Feature matching		**76.1±0.4**	**0.659±0.006**	75.7±0.4	**80.2±0.4**	**0.715±0.006**	79.7±0.4
10	Fine-tuning	(20)	76.0±0.4	**0.659±0.006**	**75.8±0.4**	80.0±0.4	0.714±0.005	**79.8±0.3**
	Fine-tuning with KL		75.7±0.4	0.652±0.006	75.2±0.4	80.0±0.4	0.709±0.006	79.2±0.5
180	Scratch	(20)	79.4±0.3	0.706±0.004	78.4±0.3	83.7±0.3	0.766±0.004	83.3±0.3
0	Direct transfer	(20)	70.1±0.6	0.561±0.008	69.4±0.7	72.8±0.6	0.592±0.009	70.6±0.7
Target domain 2: Surrey - cEEGrid dataset								
	Feature matching		**63.9±1.5**	**0.478±0.021**	**60.2±1.8**	**66.9±1.7**	**0.526±0.023**	62.3±2.0
2	Fine-tuning	(60)	61.5±1.4	0.444±0.021	58.2±1.7	64.1±1.6	0.513±0.019	61.6±1.8
	Fine-tuning with KL		62.6±1.4	0.460±0.021	59.5±1.8	66.0±1.2	0.514±0.017	**63.7±1.2**

Table 3.3: Sleep staging performance of ARNN and SeqSleepNet on source and target domains - continued

#	Method	(N)	ARNN			SeqSleepNet		
			Acc	κ	wF1	Acc	κ	wF1
	Feature matching		**68.4±2.5**	**0.543±0.036**	**64.8±2.9**	**70.4±2.6**	0.577±0.034	67.3±2.9
5	Fine-tuning	(24)	66.7±2.2	0.521±0.032	63.7±2.4	69.7±2.5	**0.584±0.033**	67.5±2.9
	Fine-tuning with KL		66.6±2.2	0.517±0.033	63.6±2.7	69.1±2.5	0.572±0.032	**67.8±2.4**
	Feature matching		**69.1±3.1**	**0.556±0.043**	65.9±3.5	71.3±3.7	**0.605±0.040**	68.6±3.6
10	Fine-tuning	(12)	68.4±2.7	0.548±0.039	**66.1±3.0**	**71.4±3.3**	0.597±0.046	**70.5±3.2**
	Fine-tuning with KL		68.1±3.1	0.542±0.045	65.7±3.8	70.6±3.2	0.577±0.047	68.9±3.3
10	Scratch	(12)	66.7±3.1	0.524±0.043	63.8±3.3	69.1±3.5	0.575±0.041	66.3±3.6
0	Direct transfer	(12)	58.2±2.3	0.410±0.031	56.7±2.4	51.4±3.7	0.375±0.032	50.4±3.1
Target domain 3: Leuven - SDS dataset								
	Feature matching		**63.1±1.0**	0.484±0.014	62.4±1.1	**67.5±1.0**	**0.544±0.014**	66.3±1.1
2	Fine-tuning	(70)	61.6±1.1	0.466±0.014	60.7±1.1	64.5±1.1	0.508±0.015	63.6±1.2
	Fine-tuning with KL		**63.1±1.0**	**0.489±0.014**	**63.0±1.1**	66.9±1.1	0.541±0.014	**66.5±1.2**
	Feature matching		**66.1±1.4**	**0.522±0.020**	**65.3±1.4**	**69.8±1.3**	**0.573±0.018**	**68.8±1.4**
5	Fine-tuning	(28)	64.6±1.4	0.504±0.020	64.0±1.4	68.2±1.3	0.552±0.019	67.8±1.4
	Fine-tuning with KL		65.0±1.5	0.512±0.020	64.9±1.5	68.1±1.5	0.554±0.021	68.2±1.6
	Feature matching		67.0±2.2	0.531±0.032	65.9±2.3	**71.3±2.2**	**0.592±0.031**	**70.4±2.2**
10	Fine-tuning	(14)	**67.1±2.1**	**0.534±0.030**	66.1±2.2	70.2±1.9	0.576±0.028	69.9±2.0
	Fine-tuning with KL		66.9±2.0	**0.534±0.028**	**66.6±2.2**	69.7±2.3	0.575±3.3	69.7±2.5
24	Scratch	(14)	67.5±2.4	0.541±0.033	65.5±2.5	73.5±2.7	0.618±0.039	70.9±2.9
0	Direct transfer	(14)	60.2±2.3	0.452±0.031	60.9±2.5	61.1±3.0	0.488±0.036	61.3±3.3

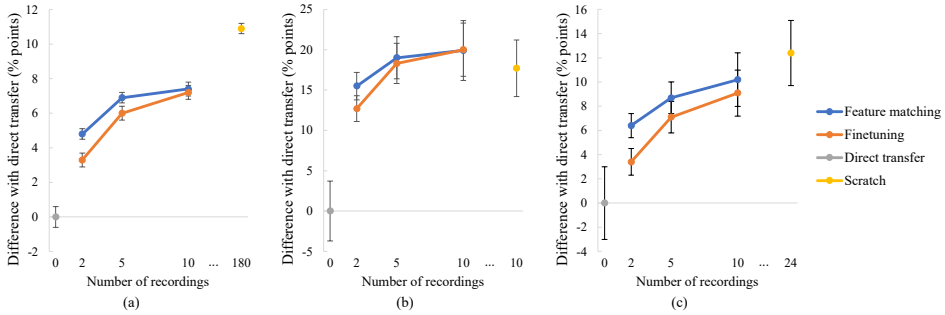


Figure 3.4: Absolute difference in accuracy of feature matching, fine-tuning and training from scratch with respect to direct transfer: a visual representation of some important results from Table 3.3. Results are shown using SeqSleepNet as a network, and for the three target datasets: (a) MASS - EOG, (b) Surrey - cEEGrid, (c) Leuven - SDS.

3.5.3 Visualizing the feature spaces

To better understand the difference between fine-tuning and feature matching, we can visualize the features learned by the sleep staging network using both techniques. For this purpose, we use the Uniform Manifold Approximation and Projection (UMAP) technique [105]. This projects the features from their high-dimensional space to two dimensions, and thus allows every sample to be plotted as a point in a 2D plane. Figure 3.5 shows the feature spaces learned by SeqSleepNet when performing transfer learning to the cross-head behind-the-ear recordings of the Leuven - SDS dataset. It plots the UMAP projections of both C4-A1 and the cross-head behind-the-ear modality. Figure 3.5(a) shows the feature space of the source modality (C4-A1) after pre-training, and Figure 3.5(b) shows how the feature space changes shape after fine-tuning on the target modality (cross-head behind-the-ear). Then, Figure 3.5(c)-(d) show the feature spaces of both modalities after feature matching. As feature matching acts on both the source modality and the target modality, we plot both the learned C4-A1 and cross-head behind-the-ear features in this case.

3.6 Discussion

In the present chapter, we propose feature matching, a novel transfer learning approach designed to transfer knowledge from a standard EEG setup to a database of a new EEG recording modality with potentially large differences

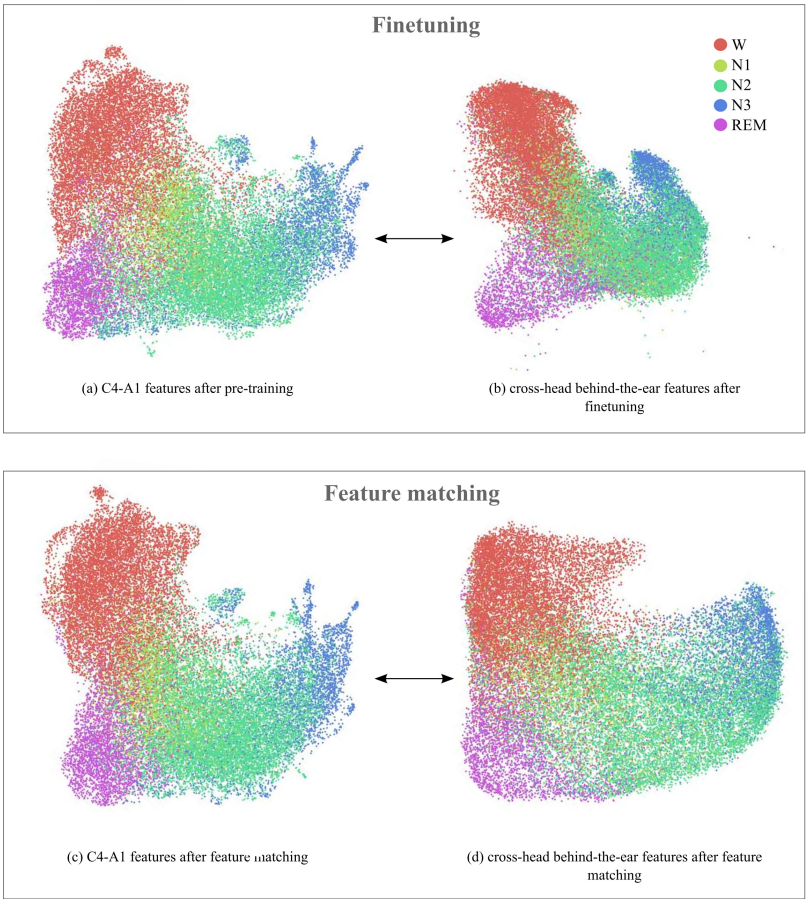


Figure 3.5: UMAP visualization of the C4-A1 and cross-head behind-the-ear modality features (both of the Leuven - SDS dataset), learned by SeqSleepNet, before and after transfer learning on 5 recordings of the Leuven - SDS dataset. Each point in a cloud represents the feature vector of one 30-second segment, and the different colors are different sleep stages. (a) C4-A1 features after pre-training on the source domain, (b) cross-head behind-the-ear features after fine-tuning on the target domain, (c) C4-A1 features after feature matching, (d) cross-head behind-the-ear features after feature matching.

in waveforms. We deploy this method in a sleep staging task, comparing its performance to fine-tuning, the state-of-the-art transfer learning approach, and to the baseline approaches of direct transfer and training from scratch. We validate our method for diverse scenarios: we use two neural network

architectures and three different target modalities of distinct datasets acquired in different sleep laboratories and recorded from different populations (Table 3.3). In addition, we analyze the effect of the transfer learning methods with varying sizes of target domain training datasets.

The need for adapting the models to the target domains is clear from the low accuracies obtained with direct transfer. All the transfer learning scenarios, even using as little as two recordings of the target modality, achieve higher accuracies than direct transfer. Certain scenarios require more adaptation than others. In cases where the performance with direct transfer is the lowest, the relative (and absolute) gains obtained from transfer learning are clearly larger. The relative percent difference in accuracy between feature matching on two recordings and direct transfer is only 6.6% for the MASS - EOG target domain, but 10.5% on the Leuven - SDS domain and 30.2% on the Surrey - cEEGrid domain, using the SeqSleepNet architecture. The EOG from the same dataset as the source domain, and therefore recorded with the same technology and from the same population, clearly requires the least adaptation. The other two target domains require more adaptation as they are recorded with different (wearable) devices, from different populations. This proves that adaptation techniques are necessary to deal with channel mismatch and other types of domain mismatch, and both the transfer learning strategies fulfill that goal.

The models are also trained from scratch on the complete target datasets. The performances achieved with this approach strongly depend on the size of the dataset. For the EOG target domain, the network trained from scratch performs better than all other methods, because it uses 180 recordings (like the baseline network trained on the source domain). For the cross-head behind-the-ear target domain, transfer learning on 10 recordings achieves a comparable performance to training from scratch on 24 recordings. For the cEEGrid target domain, training from scratch on 10 recordings performs worse than transfer learning on 10 recordings and comparably to transfer learning with 5 recordings. The amount of training data in the cEEGrid dataset, 10 recordings, is thus clearly too small to train the large amount of model parameters and achieve good generalization without relying on transfer learning. These results again show the usefulness of transfer learning in small data regimes, as it requires less data to achieve similar performances to training from scratch.

When we compare the two transfer learning techniques, feature matching and fine-tuning, the main difference between the methods lies in the minimization of the distance between source features and target features of the same samples. Figure 3.5 demonstrates the effect this has on the learned features for both modalities. Feature matching aligns the feature spaces of the two modalities (Figure 3.5(c) and (d)), whereas fine-tuning adapts the feature space of the target modality (Figure 3.5(b)) without aiming to match the source modality.

We see this effect in the similarity between the UMAP projections of the source and target features: after fine-tuning, the feature space of the source and target modality look distinctly different (Figure 3.5(a) and (b)). After feature matching, the two feature spaces are more similar in shape (Figure 3.5(c) and (d)).

In terms of classification performance, feature matching clearly has an advantage over fine-tuning in the smaller data regimes (the training scenarios on 2 and 5 recordings). When 10 (or more) recordings of the target domain are used, the two methods perform on par with each other. The less data are available from the target domain, the more the additional information from the source domain helps the model to achieve a better performance. We can understand this advantage as follows. First, a position in the feature space contains more precise information than a sleep label. The label only tells the network which region in the feature space the feature vector belongs to, whereas the use of the source feature vector adds the precise location in that feature space. Second, the feature matching technique allows to remember what was learned from the source domain, and matches the features of the target domain onto features of the source domain. As the source network is trained on a much larger dataset, the source features are of superior quality, with a better separation of the sleep stages than the target features. Aligning the target features onto those superior source features thus aids the target network to achieve a better performance as it acts as a form of regularization and implicitly exploits the information extracted through training on a much larger dataset.

We add an extra baseline method of fine-tuning with KL-divergence regularization to explore whether strong regularization in the fine-tuning approach could partly bridge the performance gap between fine-tuning and feature matching. Indeed, the fine-tuning approach with KL-divergence regularization generally performs better than the basic fine-tuning approach in the training scenario of 2 recordings, but it still mostly achieves lower performances than feature matching. This result supports the notion that feature matching adds more value than just regularization.

Our implementation of feature matching makes use of the MSE loss between paired samples, exploiting the availability of simultaneous recordings of the source and target modality. In other application areas where simultaneous recordings of a source domain and target domain might not be available, we can adapt the technique to align both domains without using the explicit correspondence between samples. The general feature matching idea and structure (Figure 3.2) does not change, but the matching loss can be implemented differently, for example with the MMD loss.

A minor disadvantage of the feature matching technique compared to fine-tuning

is the longer training time. As the feature matching structure consists of two networks instead of one, and requires training both of those networks with data from two modalities and datasets instead of one, it requires more computational power and has a longer training time (about 8 to 9 times longer in our set of experiments) than the fine-tuning technique.

3.7 Conclusion

This chapter presents feature matching, a novel transfer learning technique for deep neural networks performing sleep staging tasks. Our method is specifically tailored towards adapting sleep staging networks from standard EEG channels to new, non-standard EEG channels. Contrary to existing domain adaptation methods for EEG, this method explicitly uses the correspondence between simultaneous recordings of a standard and a non-standard EEG channel to improve the sleep staging performance on small wearable datasets. As such, in small data regimes, feature matching significantly outperforms fine-tuning, the standard transfer learning technique for this application. We conclude that this feature matching method has a lot of promise to improve sleep staging performances in small datasets with non-standard EEG modalities.

4 | Unsupervised and semi-supervised adversarial domain adaptation

This chapter is largely based on **E. R. M. Heremans**, H. Phan, P. Borzée, B. Buyse, D. Testelmans, and M. De Vos, "From unsupervised to semi-supervised adversarial domain adaptation in electroencephalography-based sleep staging," *Journal of Neural Engineering*, vol. 19, no. 4, pp. 036044, 2022.

ABSTRACT | The availability of labeled data from wearable EEG devices is very limited, yet acquiring unlabeled data is inexpensive. While supervised transfer learning techniques, like our feature matching method from Chapter 3, are useful for training sleep staging models on small labeled datasets, they can not leverage the potential unlabeled data. This chapter explores leveraging unlabeled data through unsupervised and semi-supervised adversarial domain adaptation on wearable sleep datasets from diseased patient populations. Different practical aspects of the adversarial domain adaptation framework are examined, including the impact of (pseudo-)labels from the target dataset and domain mismatch between the source and target data. The method is also implemented for personalized patient-specific models. Results demonstrate the effectiveness of adversarial domain adaptation in sleep staging on wearable data, with relative accuracy gains of 7%-27%. The performance is further improved by adding target domain (pseudo-)labels and choosing an appropriate source dataset. Furthermore, unsupervised personalization improves performance by 1%-2%. In conclusion, adversarial domain adaptation provides a flexible framework for semi-supervised and unsupervised transfer learning, facilitating accurate sleep staging on EEG acquired from wearable devices.

4.1 Introduction

While the availability of labeled data from wearable EEG devices is very limited, acquiring unlabeled data is relatively inexpensive. Current wearable datasets remain small, but the widespread adoption of wearable technology for large-scale sleep monitoring is expected to sharply increase the volume of available unlabeled data. While supervised transfer learning techniques, like the one presented in the previous chapter, excel at training sleep staging models on small labeled datasets, they do not utilize the potential of this vast, future pool of unlabeled data.

Exploiting this resource requires unsupervised techniques that rely solely on unlabeled wearable EEG data or semi-supervised techniques that leverage both unlabeled and labeled data. Very recently, unsupervised domain adaptation techniques have found their way to EEG-based classification tasks [106–110]. Most of these techniques are based on domain-invariant feature learning, either by domain-adversarial training of neural networks [111], or using the MMD loss [103, 104]. A few studies focus on adversarial domain adaptation for sleep staging specifically [95–97]. These methods successfully cope with domain mismatch between different sleep databases without requiring any labels from the target domain. All these studies exploit a common framework with some minor variations. The methods are demonstrated on traditional PSG sleep databases, with minimal differences between recorded channels of the source datasets and target datasets.

Building upon these previous works, this chapter introduces several contributions. First, instead of creating a novel domain adaptation method, we assess the performance of the common adversarial domain adaptation backbone in novel, real-world use cases. The target domains are three different databases with various non-traditional EEG derivations and real wearable EEG data. The data are acquired from different populations and include a realistic use case of elderly and diseased patients. Second, we investigate the added value of using pseudo-labels or real labels of the target domain, comparing a semi-supervised approach to fully unsupervised adversarial domain adaptation. To our best knowledge, the use of target labels was not yet investigated in domain adaptation studies for sleep staging, and the individual impact of pseudo-labels was not yet discussed. We hypothesize that a limited number of labels may help overcome the large mismatch between traditional PSG and wearable EEG datasets. Therefore, we augment the unsupervised domain adaptation

framework used in previous studies to a semi-supervised framework allowing for labeled target data. We also add a separate target domain classifier to allow for more flexibility in this framework. Third, we evaluate the effect of the different causes for domain mismatch between a source domain and a target domain. To do so, we investigate alternative source domains to match with the target domains. Lastly, we demonstrate that the same framework can be used for personalization, applying the adversarial domain adaptation method to single subjects.

The remainder of this chapter is structured as follows. Section 4.2 starts with an introduction to domain adaptation and transfer learning and explains the proposed adversarial domain adaptation framework with its different variations. Section 4.3 describes the datasets and neural network architecture used to evaluate the methods. Section 4.4 discusses the experiments conducted to assess the performance of the transfer learning methods in different scenarios. Section 4.5 then shows the obtained results, which are discussed in Section 4.6. Section 4.7 extracts some final conclusions and recommendations from this chapter.

4.2 Methods

4.2.1 Domain adaptation

Transfer learning was formally defined in Section 3.3.1. Domain adaptation is a common type of transfer learning in which the task remains unchanged ($\mathcal{T} = \mathcal{T}_S = \mathcal{T}_T$), but the source domain differs from the target domain ($\mathcal{D}_S \neq \mathcal{D}_T$). In domain adaptation, labeled source data are usually assumed available. As such, we call a domain adaptation technique unsupervised, semi-supervised, or supervised depending on the use of labels from the target domain [83].

In the application to automated sleep staging on wearable EEG recordings, we select a standard EEG channel in a large public PSG database as the source domain, $\mathcal{D}_S$. Knowledge is transferred from this domain to a target domain, $\mathcal{D}_T$, consisting of a small dataset of non-standard EEG recordings. The main difference between the source and target dataset is in the domains, owing to the difference in electrode positions. The task is not profoundly changed between both datasets, as it always consists of classifying 30-second data segments into five sleep stages. As a result, this can be regarded predominantly as a domain adaptation problem. However, when the patient population or the annotator changes between the datasets, the predictive function $f(.)$ between the input data and output labels changes as well. Hence, the tasks $\mathcal{T}_S$ and $\mathcal{T}_T$ are not

actually identical, so we argue this should be regarded as a general transfer learning problem. When target and source tasks are different, this, in turn, implies that some labeled target data are required to infer the target predictive function $f_T(.)$, as argued by theoretical works [83, 102]. Therefore, while some of the discrepancies between the source and target can be overcome by using unlabeled target data, the performance on the target domain will be further improved with the addition of target domain labels.

4.2.2 Adversarial domain adaptation

In adversarial domain adaptation, the domain discrepancy between a source domain and a target domain is overcome through two competing objectives. A feature extractor F extracts features from the input data. These features are then processed by a classifier C predicting the class labels and by a domain discriminator D predicting whether samples are from the source domain or the target domain. The feature extractor tries to confuse the domain discriminator by producing domain-invariant features and works with the classifier to produce features that allow distinguishing different classes. Figure 4.1 shows this idea in a schematic representation. In the application of sleep staging, F is the sleep staging network, and C can simply be its classification layer. The features computed by the one-before-last layer of the network are thus fed to both the classification layer and domain discriminator.

Typically, in adversarial domain adaptation, the feature extractor and classifier are both shared between the source and the target domain ($F \equiv F_C \equiv F_T$, $C \equiv C_S \equiv C_T$). While a shared F is not a necessity but rather a common design choice [112], we opt to adopt this approach in alignment with previous studies by Nasiri [95] and Yoo [97]. A shared C is necessary when there is no loss function to optimize a target classifier C_T independently. This is the case when no target domain (pseudo-)labels are used. When we do use (pseudo-)labels of the target domain, we can train an independent target classifier to learn the target task independently ($\mathcal{T}_S \neq \mathcal{T}_T$, see Section 4.2.1).

Three versions of the adversarial domain adaptation technique are implemented: a basic method, a method with pseudo-labels, and a semi-supervised method with some labels from the target domain. In the pseudo-labels method and target labels method, a separate target classifier C_T is used. The following sections explain these three different adversarial domain adaptation methods in more detail, with a summary in Table 4.1. Source code is available at https://github.com/elisabethRMH/adversarial_DA.

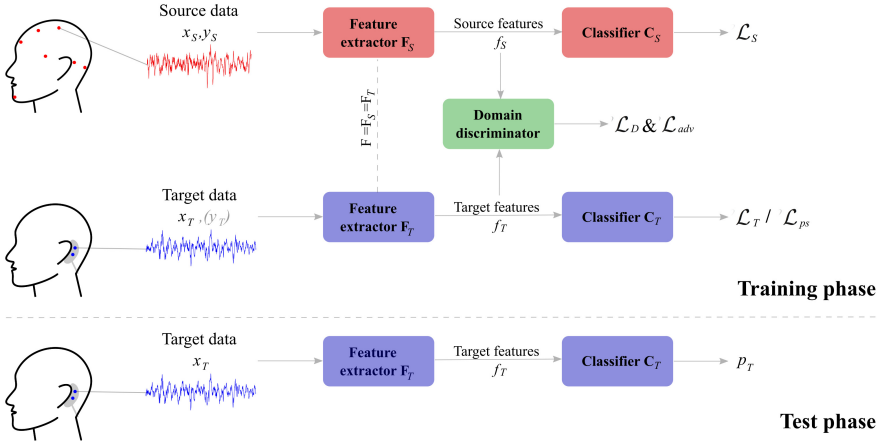


Figure 4.1: Schematic illustration of adversarial domain adaptation. Subscripts S and T stand for source and target domain, respectively. The feature extractors are always shared ($F_S \equiv F_T \equiv F$). The classifiers are shared in the basic method ($C_S \equiv C_T \equiv C$), but not in the pseudo-labels and target labels method. The loss function $\mathcal{L}_{ps}$ is used in the pseudo-labels method, $\mathcal{L}_T$ is used in the target labels method.

4.2.3 Basic method

The basic adversarial domain adaptation method is implemented as follows. The classifier and feature extractor are trained for the classification in the source domain using the standard categorical cross-entropy loss:

$$\mathcal{L}_S = -\mathbb{E}_{x_S \sim P_S} \sum_{c=1}^{N_c} y_S^c \log\, p_S^c, \tag{4.1}$$

where $p_S = C_S(F(x_S)) = C(F(x_S))$ is the class probability output of the network for a source sample, and y_S is the one-hot sleep stage label. c is the class number and N_c the number of classes. The domain discriminator is trained with the standard binary cross-entropy loss to classify the domain of each sample:

$$\mathcal{L}_D = -\mathbb{E}_{x_S \sim P_S} \log\,(1 - d'_S) - \mathbb{E}_{x_T \sim P_T} \log\, d'_T, \tag{4.2}$$

where $d'_T = D(F(x_T))$ and $d'_S = D(F(x_S))$. The feature extractor F is trained with the adversarial loss to confuse the domain discriminator. The adversarial loss function corresponds to $\mathcal{L}_D$ with inverted labels:

$$\mathcal{L}_{adv} = -\mathbb{E}_{x_S \sim P_S} \log\ d'_S - \mathbb{E}_{x_T \sim P_T} \log\,(1 - d'_T). \tag{4.3}$$

The basic adversarial domain adaptation framework is governed by the following end-to-end optimization:

$$\min_{D}\ \mathcal{L}_D, \tag{4.4}$$

$$\min_{F,C}\ \mathcal{L}_S + \lambda_{adv}\mathcal{L}_{adv}. \tag{4.5}$$

4.2.4 Pseudo-labels

Prior work on adversarial domain adaptation for EEG applications has already used multiple additions and tricks to improve upon the basic framework [95–97]. Pseudo-labels [96] and entropy minimization [95, 97] are techniques aimed at mitigating the lack of labels of the target domain. Both approaches include the addition of a categorical cross-entropy loss based on the target samples:

$$\mathcal{L}_{ps} = -\mathbb{E}_{x_T \sim P_T} \sum_{c=1}^{N_c} \tilde{y}_T^c \log\ p_T^c, \tag{4.6}$$

in which p_T is the class probability output of the network for a target sample, and $\tilde{y}_T^c$ is a substitute for the true label of this target sample. In entropy minimization, the true label is simulated by the classifier's output itself ($\tilde{y}_T^c = p_T^c$). Pseudo-labels are used as a more general term for an approximation of the true labels.

In this chapter, pseudo-labels are implemented as follows. Different from previous works on adversarial domain adaptation for sleep staging, a separate classifier is defined for the source domain (C_S) and the target domain (C_T). The source domain classifier is trained with $\mathcal{L}_S$ as defined in the loss function (4.1). The target domain classifier is trained with the loss function (4.6), where $\tilde{y}_T = C_S(F(x_T))$ is the probability output of the source classifier for the target samples, and $p_T = C_T(F(x_T))$ is the probability output of the target classifier for the target samples. This way, $\mathcal{L}_{ps}$ prevents the target classifier from straying too far from the source classifier while also favoring outputs with minimal entropy, i.e. with a higher probability for one class.

In conclusion, the pseudo-labels version of adversarial domain adaptation is governed by the following optimization procedure:

$$\min_{D} \ \mathcal{L}_D, \tag{4.7}$$

$$\min_{F,C_S,C_T} \ \mathcal{L}_S + \lambda_{adv}\mathcal{L}_{adv} + \lambda_{ps}\mathcal{L}_{ps}. \tag{4.8}$$

4.2.5 Target labels

As explained in Section 4.2.1, the sleep staging task itself can differ between a source and target dataset. Learning the target task then requires some target labels. We thus investigate how much the target domain performance can be boosted by adding a limited amount of target labels. This semi-supervised approach is implemented by using labels of only two recordings of the target domain. C_T is trained with real target domain labels of a limited number of samples with the categorical cross-entropy loss function:

$$\mathcal{L}_T = -\mathbb{E}_{x_T \sim P_T} \sum_{c=1}^{N_c} y_T^c \log \ p_T^c. \tag{4.9}$$

This results in the following optimization framework for the adversarial domain adaptation with target labels:

$$\min_{D} \ \mathcal{L}_D, \tag{4.10}$$

$$\min_{F,C_T,C_S} \ \mathcal{L}_S + \lambda_{adv}\mathcal{L}_{adv} + \mathcal{L}_T. \tag{4.11}$$

Table 4.1: The three different adversarial domain adaptation methods.

	Classifiers	Loss functions	Hyperparameters
Basic	Shared (C)	$\mathcal{L}_D, \mathcal{L}_S, \mathcal{L}_{adv}$	$\lambda_{adv} = 0.01$
Pseudo-labels	Unshared ($C_S \neq C_T$)	$\mathcal{L}_D, \mathcal{L}_S, \mathcal{L}_{adv}, \mathcal{L}_{ps}$	$\lambda_{adv} = 0.01, \lambda_{ps} = 0.01$
Target labels	Unshared ($C_S \neq C_T$)	$\mathcal{L}_D, \mathcal{L}_S, \mathcal{L}_{adv}, \mathcal{L}_T$	$\lambda_{adv} = 0.01$

4.3 Materials

4.3.1 Source domain: PSG datasets

MASS – C4 An appropriate choice for the source domain in transfer learning experiments is a standard EEG channel of a large public PSG dataset. In all experiments but one, we use the 200 recordings of MASS [65] (Section 2.1.1) for this purpose. The C4-A1 channel of this database is selected as the source domain for all experiments in this chapter unless otherwise mentioned.

Leuven – C4 In one experiment, a different source database is used to show the influence of changing the source domain on the transfer learning process. This source database is extracted from a large PSG dataset, consisting of 218 recordings of patients with suspected sleep apnea [113]. The dataset was recorded at the sleep laboratory of the University Hospitals Leuven and annotated according to AASM standards [67]. From this PSG dataset, we select the C4-A1 channel of 38 recordings corresponding to patients over 60 years of age. This dataset is further referred to as the Leuven - C4 database.

4.3.2 Target domain: wearable EEG datasets

Three different target domains are used to investigate the transfer learning scenarios in various setups reflecting possible use cases and wearable EEG configurations.

Surrey – cEEGrid The first target domain is the Surrey – cEEGrid dataset [35, 68] (see Section 2.1.1), recorded from 12 participants with the wearable cEEGrid device. In this chapter, the right-ear front-versus-back cEEGrid derivation is used.

Dreem – Headband The second target dataset is the Dreem – Headband dataset [36] (see Section 2.1.1), recorded with the wearable headband from 25 subjects. In the present chapter, the F7-F8 derivation of the Dreem Headband is used as a target domain.

Leuven – SensorDot Senior Third, the first 46 recordings from the Leuven - SDS sleep dataset are used (see Section 2.1.1). As in Chapter 3, we simulate

a wearable behind-the-ear EEG by using the cross-head behind-the-ear EEG channel.

4.3.3 Sleep staging network

Adversarial domain adaptation can be applied to any sleep staging network architecture, as it only requires taking features learned by this network and feeding these to a domain discriminator. Both recurrent and convolutional architectures were successfully used in combination with adversarial domain adaptation [97]. In order to validate the proposed methods, the SeqSleepNet architecture [46] (see Section 2.2) was used to extract relevant features from the EEG signals.

4.4 Experiments

A number of experiments were performed to investigate multiple factors influencing the performance of adversarial domain adaptation. In each of these experiments, the feature extractor and classifier were first initialized with weights pre-trained on the source domain. Then, one of the adversarial domain adaptation methods from Section 4.2 was applied. Source and target domain samples were fed to the feature extractor, and the feature extractor was trained to match the two domains. The classification performance on the target domain was investigated.

4.4.1 Experimental setup

Training parameters

The base network was pre-trained for 10 training epochs on all 200 C4-A1 recordings of the MASS dataset, except for 10 recordings used as a validation set to retain the best model. Then, for every domain adaptation experiment, cross-validation was performed on the target dataset to obtain average performance values. For every fold of a cross-validation experiment, the target dataset was divided into a training set, a validation set, and a test set. The training set was used to train the adaptation method for 20 training epochs, and the validation set was used to evaluate the model after every 100 training steps and retain the best-performing one. The best model was then evaluated on the test set. The test performances were averaged over all the cross-validation folds. For the

Surrey - cEEGrid dataset consisting of 12 recordings, 12-fold cross-validation was performed. For the Dreem - Headband dataset, 12-fold cross-validation was performed on the 25 recordings. The Leuven - SDS dataset of 46 recordings was split through 23-fold cross-validation.

Minibatch construction

During training, minibatches were constructed with

1. labeled data from the source domain,
2. unlabeled data from the target domain,
3. (only in the semi-supervised experiments:) labeled data from the target domain.

In every minibatch, the number of labeled samples from the target domain was fixed to 8 in the semi-supervised experiments. The number of unlabeled samples of the target domain was such that one training epoch would correspond to one pass through both the labeled and unlabeled part of the dataset. The amount of data from the source domain was balanced with the amount of target data.

Performance metrics

After applying the adversarial domain adaptation training procedure, the final model was evaluated on an independent test set of the target domain as $p_T = C_T(F(x_T))$. This result was compared to the ground truth target labels y_T using the classification accuracy, κ, and wF1.

4.4.2 Influence of target (pseudo-)labels

In the first set of experiments, we aimed to quantify the effect of adversarial domain adaptation on the wearable EEG target datasets and determine the impact of pseudo-labels and real target labels on the performance. Various baseline methods were implemented in order to evaluate the three adversarial domain adaptation methods defined in Section 4.2. The lower limit baseline was defined as the performance obtained by directly applying the network to the target domain after pre-training on the MASS C4-A1 dataset ('direct transfer'). The model was also trained on the complete labeled target dataset, with

supervised training from scratch and supervised fine-tuning of the pre-trained network. The latter method represents the upper limit to the performance.

These experiments were performed for all three target domains: the Surrey - cEEGrid dataset, the Headband data of the Dreem dataset, and the cross-head behind-the-ear modality of the Leuven - SDS dataset. The Surrey - cEEGrid dataset and Dreem - Headband dataset both allow to investigate transfer learning on real wearable data acquired at different sleep laboratories from the source data of the MASS dataset. While the cEEGrid data are recorded from behind the ear, the Headband data are recorded with forehead electrodes. The Leuven - SDS target domain allows to analyze transfer learning on a new EEG modality acquired at another sleep laboratory and from a widely different population. The diseased and elderly patient population of this dataset reflects the real use case for wearable sleep monitoring and allows us to validate our sleep staging methods on the target population with suspected sleep-wake disturbances.

4.4.3 Influence of amount of domain mismatch

When a target and source domain are more similar, the domain mismatch is smaller, and the domain adaptation problem gets easier. The same holds for the difference in the source and target task in transfer learning. In each of the two following experiments, we selected a different source dataset with a closer resemblance to a target dataset. We tested whether such a better-matched source domain improved the performance after transfer learning to the target domain. Two different sources of mismatch were investigated in this way. To ensure a fair comparison, the source domain was only changed in the adversarial domain adaptation training, so pre-training was performed on the original source domain (C4-A1 of MASS) in all experiments.

Influence of recording setup

The first experiment tested the influence of the mismatch between the recording setup of the source and target domain. The F7-F8 forehead derivation of the Dreem dataset greatly differs from the C4-A1 derivation of the MASS dataset. However, the MASS dataset's EOG signal (EOG left-right) should resemble the F7-F8 derivation much more. For that reason, we tested adversarial domain adaptation with MASS's EOG derivation as a source domain and the Dreem dataset's F7-F8 derivation as a target domain. The obtained performance was compared to the performance obtained with MASS's C4-A1 channel as a source domain.

Influence of other factors

The second experiment tested the impact of the mismatch between the population, recording equipment, and scorer of the source data and target data. The Leuven - C4 dataset was specifically selected as an almost perfect match for the Leuven - SDS dataset with regard to all three factors. Indeed, both datasets are recorded from elderly suspected sleep apnea patients in the same sleep laboratory and hence with the same equipment, and even scored by the same scorer. The sleep staging model was thus adapted to the Leuven - SDS dataset (right ear - A1 derivation) with two source domains: the highly matched Leuven - C4 dataset and the general MASS - C4 dataset. The source EEG derivation was C4-A1 in both cases.

4.4.4 Personalization

Adversarial domain adaptation could also provide an elegant way to personalize a sleep staging network to an individual subject. A last set of experiments was conducted to determine the effect of unsupervised personalization, using adversarial domain adaptation with pseudo-labels. Since only one recording was available per subject, these experiments required some adjustments of the experimental setup defined for the other experiments.

Regarding the pre-training, it made sense to start the personalization process for each subject with a network adapted to the relevant EEG modality and dataset. The feature extractor and classifier weights were thus initialized with network weights pre-trained with supervised fine-tuning on the relevant dataset and EEG derivation (Surrey - cEEGrid, Dreem - Headband, or Leuven - SDS). Following the cross-validation scheme from Section 4.4.1, in every fold, the base network was first fine-tuned to other recordings of the dataset (the training set of the fold). Then, adversarial domain adaptation was separately performed on each recording in the fold's test set. The new 'source dataset' in this personalization process consisted of the training set of the current fold, and the 'target dataset' was one of the test set recordings. To avoid confusion with the source and target datasets as defined for the other experiments, these are further referred to as the training set and test recording in the context of personalization.

Pre-training was achieved with transfer learning to the other recordings of the target dataset, and was thus performed with the same experimental setup as the transfer learning experiments (see Section 4.4.1). In the personalization step, the network was trained for 10 training epochs instead of 20. Instead of selecting the best-performing model using a validation set, we retained the model with the lowest pseudo-label loss $\mathcal{L}_{ps}$ for evaluation. This change in the experimental

setup was necessary because we could not use an independent labeled validation set as the target domain was just one recording. Minibatches consisted of 32 samples of the training set and the test recording. After performing the cross-validation as described, results were averaged over all the test recordings.

The personalization experiments were carried out on all three target datasets: the Surrey - cEEGrid dataset, the Dreem - Headband dataset, and the Leuven - SDS dataset. Adversarial domain adaptation for personalization was performed with the pseudo-labels method and compared with two baselines. The first baseline was the performance of the pre-trained baseline network, in this case, trained with supervised fine-tuning to the relevant target dataset. The second baseline was the result obtained with a simple adaptation of the normalization statistics in the batch norm layers of the sleep staging network. This commonly known domain adaptation technique [83] was recently demonstrated in personalized sleep staging [114]. The method adjusts normalization statistics to a target domain without training any network weights. By comparing against this baseline, we investigated whether adversarial domain adaptation accomplishes more than only adapting these normalization statistics.

4.5 Results

4.5.1 Influence of target (pseudo-)labels

Figure 4.2 shows the influence of adversarial domain adaptation on the sleep staging performance in the target domain. The impact of using (pseudo-)labels of the target domain is also shown. The accuracy is plotted for all three proposed adversarial domain adaptation methods: the basic method (ADA-0), the pseudo-labels method (ADA-ps), and the method with real labels of two recordings (ADA-2). The performance of these adversarial domain adaptation methods is compared to the three baselines: direct transfer of the pre-trained base network (DT), supervised training from scratch (FS), and fine-tuning on the target dataset (FT). Table 4.2 shows the complete results, with all three performance metrics. The performance metrics are represented as their mean and standard error over all cross-validation folds.

The results clearly show that all three adversarial domain adaptation methods achieve performances between the direct transfer and training from scratch baselines. The supervised fine-tuning baseline performs better than training from scratch. Overall, out of the three adversarial domain adaptation methods, the version with true target labels of two recordings (ADA-2) performs best,

followed by the pseudo-label version (ADA-ps), and lastly the basic method (ADA-0).

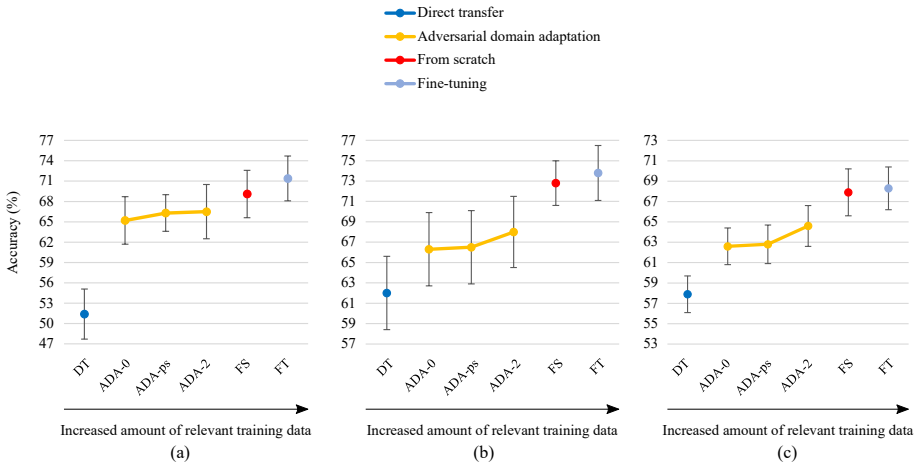


Figure 4.2: The accuracy of the three adversarial domain adaptation techniques on the three target datasets. Direct transfer (DT), fully supervised training (FS), and supervised transfer learning (TL) are shown as baselines. ADA designates adversarial domain adaptation. ADA-0 is the basic ADA, ADA-ps is the pseudo-label version, and ADA-2 is the ADA with labels of two target recordings. From left to right, there is an increased amount of relevant training data. The error bars indicate the standard error over the cross-validation folds. (a) The Surrey - cEEGrid dataset. (b) The Dreem - Headband dataset. (c) The Leuven - SDS dataset.

The hyperparameters λ_{adv} and λ_{ps} define the relative weight of the adversarial domain loss function and the pseudo-label loss function during training. Figure 4.3 shows a sensitivity analysis to these hyperparameters on the Surrey - cEEGrid dataset. As can be seen, the hyperparameter values $\lambda_{adv} = 0.01$ and $\lambda_{ps} = 0.01$ are optimal for this dataset. We selected these values, applying $\lambda_{adv} = 0.01$ and $\lambda_{ps} = 0.01$ for all three datasets. The trends and results were consistent across the three datasets without overfitting to these hyperparameters.

4.5.2 Influence of domain mismatch

Figure 4.4 shows the effect of different sources of domain mismatch on the adversarial domain adaptation results. The figure shows the mean and standard error of the sleep staging accuracy for different training methods: adversarial domain adaptation with no target labels (ADA-0), with pseudo-labels (ADA-ps) and with labels of two recordings (ADA-2).

Table 4.2: The results of the three adversarial domain adaptation techniques: the basic version using no labels of the target domain, the pseudo-labels version and the version with target domain labels of two recordings. The compared baselines are direct transfer with no retraining, fully supervised training from scratch on the target dataset, and supervised fine-tuning on the target dataset. The sleep staging performance is reported for the three target domains, as mean ± standard error over all the cross-validation folds. Metrics are the accuracy (acc), κ and wF1.

Surrey - cEEGrid				
Method		Acc	κ	wF1
Direct transfer		51.4±3.7	0.375±0.032	50.4±3.1
ADA	Basic	65.2±3.5	0.512±0.046	62.4±3.3
ADA	Pseudo-labels	66.3±2.7	0.517±0.039	63.2±2.8
ADA	Target labels	66.5±4.0	0.527±0.055	64.3±4.0
From scratch		69.1±3.5	0.575±0.041	66.3±3.6
Fine-tuning		71.4±3.3	0.597±0.046	70.5±3.2
Dreem - Headband				
Method		Acc	κ	wF1
Direct transfer		62.0±3.6	0.460±0.045	61.4±3.6
ADA	Basic	66.3±3.6	0.512±0.049	66.0±3.6
ADA	Pseudo-labels	66.5±3.6	0.517±0.047	66.3±3.5
ADA	Target labels	68.0±3.5	0.537±0.048	67.8±3.5
From scratch		72.8±2.2	0.599±0.030	71.1±2.1
Fine-tuning		73.8±2.7	0.619±0.035	73.5±2.6
Leuven - SDS				
Method		Acc	κ	wF1
Direct transfer		57.9±1.8	0.440±0.024	58.4±1.8
ADA	Basic	62.6±1.8	0.482±0.024	63.1±1.8
ADA	Pseudo-labels	62.8±1.9	0.485±0.025	63.2±2.0
ADA	Target labels	64.6±2.0	0.506±0.028	64.3±2.2
From scratch		67.9±2.3	0.538±0.032	65.2±2.4
Fine-tuning		68.3±2.1	0.550±0.030	66.9±2.2

Figure 4.4(a) displays the influence of the mismatch between recording setups on the classification performance on the Dreem - Headband dataset. There is a clear difference in performance between the scenario with the EOG channel and the C4-A1 channel as the source domain. With the EOG channel as a source domain, the mean accuracy is higher for all three adversarial domain adaptation methods. Figure 4.4(b) shows the impact of the mismatch between the population, scorer, and equipment on the classification performance in the Leuven - SDS dataset. The sleep staging accuracy on this target dataset

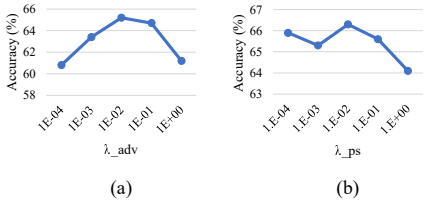


Figure 4.3: Sensitivity analysis to the hyperparameters λ_{adv} and λ_{ps} in Surrey - cEEGrid dataset. (a) λ_{adv} in the basic adversarial domain adaptation experiment ($\lambda^*_{adv} = 0.01$), (b) λ_{ps} in the pseudo-label experiment ($\lambda^*_{ps} = 0.01$).

is higher with the Leuven - C4 dataset as a source domain than with the MASS dataset as a source domain. Again, this is true for all three adversarial domain adaptation methods (see the green line compared to the yellow line in Figure 4.4(b)).

These results clearly indicate that the mismatch between the source and target dataset influences the obtained classification performance on the target dataset (for a further discussion, see Section 4.6).

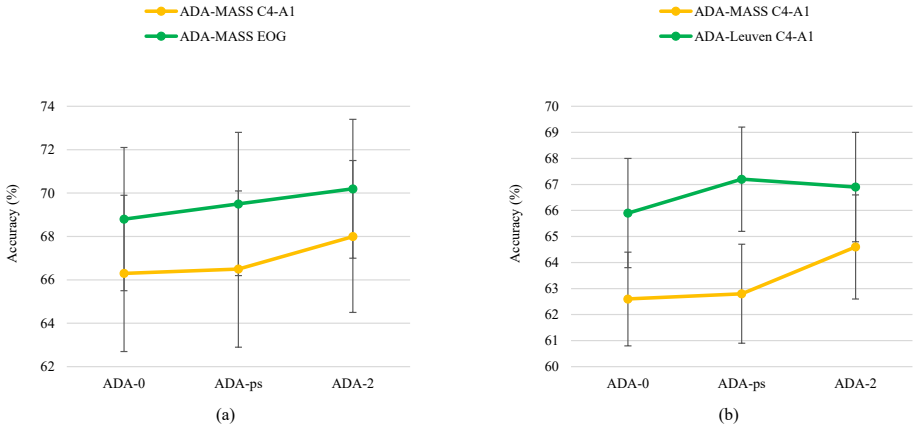


Figure 4.4: The influence of the domain mismatch between the source domain and target domain. ADA designates adversarial domain adaptation. ADA-0 is the basic ADA, ADA-ps is the pseudo-label version, and ADA-2 is the ADA with labels of two target recordings. The error bars indicate the standard error over the cross-validation folds. (a) For the Dreem - Headband dataset, the MASS dataset's EOG is compared to the MASS dataset's C4-A1 derivation as a source domain. (b) For the Leuven - SDS dataset, the Leuven - C4 dataset is compared to the MASS dataset as a source domain.

4.5.3 Personalization

Table 4.3 shows the effect of personalization on individual patients. The performance before personalization is shown (non-pers), as well as the performance after personalization with the batch norm method (BN pers) and with adversarial domain adaptation (ADA pers). For each metric and target dataset, the table reports the mean and standard error over all the patients in this target dataset.

In almost all experiments, both personalization methods improve the performance compared to the baseline of fine-tuning to other patients of the same dataset. It is also observed that adversarial domain adaptation outperforms or achieves similar results to batch norm personalization in all datasets.

Table 4.3: Results of personalization on the three target domains. Mean ± standard error of the sleep staging performance over all the subjects after transfer learning to the dataset without personalization (Non-pers), batch normalization personalization (BN pers), and adversarial domain adaptation personalization (ADA pers). Accuracy (acc), κ, and wF1 are shown. For each metric, the highest mean is shown in bold.

Surrey - cEEGrid			
	Acc	κ	wF1
Non-pers	71.4±3.3	0.597±0.046	70.5±3.2
BN pers	72.5±3.4	0.613±0.047	72.4±2.9
ADA pers	**72.8±3.6**	**0.618±0.049**	**72.7±3.2**
Dreem - Headband			
	Acc	κ	wF1
Non-pers	73.6±2.7	0.625±0.030	73.8±2.6
BN pers	72.9±2.5	0.610±0.032	73.1±2.4
ADA pers	**74.5±2.7**	**0.638±0.030**	**74.6±2.6**
Leuven - SDS			
	Acc	κ	wF1
Non-pers	68.4±1.8	0.544±0.026	67.4±1.8
BN pers	**69.1±1.7**	0.551±0.026	**68.4±1.6**
ADA pers	69.0±1.8	**0.558±0.025**	68.2±1.7

4.6 Discussion

This chapter proposes a unified adversarial domain adaptation framework for real-world use cases and evaluates its performance in the context of sleep staging on wearable EEG recordings. Experiments were performed to investigate the influence of key elements in this framework: the use of pseudo-labels and real labels from the target domain, the similarity of the source domain to the target domain, and the potential for personalization.

First, the adversarial domain adaptation strategy clearly achieves its purpose in real applications of wearable EEG recordings and diseased patient populations. The basic adversarial domain adaptation achieved relative accuracy improvements of 8% to 27% compared to the direct transfer scenario (see Table 4.2). These gain margins are higher than those reported for the SeqSleepNet network by Yoo et al. [97], ranging from -1% to 13%. Our higher margins of improvement can be directly related to the increased domain mismatch in our experiments. In [97], the source and target domains consisted of similar, mostly healthy populations with comparable EEG channel configurations. In our datasets, there was a large mismatch between the populations and channel configurations. This resulted in lower baseline accuracies for the direct transfer scenario and larger gain margins on average.

Second, this chapter generalizes adversarial domain adaptation to a semi-supervised framework for EEG-based classification and sleep staging applications. The addition of a separate target domain classifier trained with pseudo-labels or real labels from the target domain clearly has a positive impact on the classification performance in the target domain. Pseudo-labels resulted in rather small improvements relative to the basic unsupervised method, with relative increases in accuracy 0% to 2%. The improvement by using real target labels led to relative accuracy gains of 2% to 3%. We conclude from these results that the adversarial domain adaptation backbone is easily generalized to a semi-supervised framework. If a limited amount of target labels is available, it is beneficial to use them. If no labels of the target domain are available, pseudo-labels can be employed as an imperfect substitute. Our results support the idea that the classification tasks in the source dataset and target dataset differ to an extent, as this could explain the large gains obtained from training a separate target domain classifier with target labels.

The domain mismatch experiments were designed to investigate the influence of the similarity between the source domain (and task) and the target domain (and task) in the transfer learning process. In both experiments, during the adversarial domain adaptation training, the original source dataset was replaced with a new source dataset, which more closely matched the target dataset. The

pre-training of the network was still performed on the original source dataset in both cases, so all adversarial domain adaptation experiments started from the same baseline network. The first domain mismatch experiment investigated the influence of the source domain channel (see Figure 4.4(a)). By using the EOG channel instead of the C4-A1 channel of the MASS dataset, the accuracy improved by a relative 3% to 5%. This clearly indicates that the match between a source channel and target channel has a large influence over the target performance. In practical applications involving wearable EEG datasets, it is sensible to select the PSG channel that most closely resembles the new wearable derivation as the source domain channel.

The second domain mismatch experiment tested the influence of other factors of the source dataset (see Figure 4.4(b)). Being a highly controlled dataset of the same population, acquired in the same sleep laboratory and scored by the same scorer, the Leuven - C4 dataset is an almost perfect match for the Leuven - SDS dataset. The only substantial difference between these two datasets is the EEG channel, which is the standard C4-A1 in the source domain and the right ear - A1 channel in the target domain. Using this matched source domain instead of the MASS source domain improved accuracy by 4% to 7%. Therefore, we conclude from this experiment that the match between the populations, scorers, and recording equipment of the source and target datasets also greatly influences the target performance. In real clinical applications, however, it may often be unrealistic to find a source dataset that closely matches the target dataset in all these aspects. Nevertheless, when selecting a source domain, any parameter that influences the similarity of the source and target domains should be considered to optimize the performance in the target domain. A second noteworthy observation from using the Leuven - C4 as a source dataset was the lack of improvement when using true labels instead of pseudo-labels (see ADA-ps and ADA-2 in Figure 4.4(b)). As the population and scorer are the same in the Leuven - SDS and Leuven - C4 datasets, this is a more 'pure' domain adaptation problem in which the sleep scoring task remains unchanged. This argument could explain why the target labels in this scenario did not add value. Again, this result supports the idea that the task mismatch between datasets is the reason why target labels are useful.

Lastly, the personalization experiments compared a baseline model fine-tuned on the relevant dataset to a model personalized to a specific recording (see Table 4.3). The adversarial personalization results were also compared to batch norm personalization as a simpler alternative method. Overall, both personalization strategies improved upon the non-personalized baseline. The adversarial strategy systematically improved the baseline by a relative 1% to 2%. For all three datasets, it outperformed or performed on par with the batch norm method, which improved the baseline in two out of three datasets.

Hence, this investigation indicates that adversarial domain adaptation can be applied for personalization. Further work is needed to determine to what extent it outperforms the simple batch norm method. The margin of improvement obtained with the adversarial domain adaptation strategy may well increase with increasing numbers of recordings of a subject. Indeed, more recordings might lead to a more accurate target distribution with sufficient representation for all sleep stages.

4.7 Conclusion

In this chapter, adversarial domain adaptation was applied to several real-world sleep staging problems. We can summarize our findings in a number of recommendations for adversarial domain adaptation in general, and some for sleep staging in particular.

First, adversarial domain adaptation is mostly implemented as an unsupervised method, but it can easily be generalized to a semi-supervised method. When some labels of the target domain are available, they should be used for improving the accuracy on the target domain. This is mostly beneficial when the source and target classification task may differ. In sleep staging tasks, the classification task may change with the study population and scorer, making semi-supervised learning with some target labels the superior option. Pseudo-labels can serve as an imperfect substitute when no true labels are available. When multiple labeled source domains are available to choose from, it is advisable to select the one that is the most similar to the target domain. For sleep staging, both the recording setup and the practical aspects of the sleep study such as the scorer, patient population, and recording equipment are of consequence. Lastly, we demonstrated that adversarial domain adaptation can elegantly achieve personalization of a model to a specific recording of an individual subject.

Part III

Uncertainty estimation to increase trust

5 | An uncertainty-guided deep learning pipeline for automated sleep staging

This chapter is largely based on **E. R. M. Heremans**, N. Seedat, B. Buyse, D. Testelmans, M. van der Schaar, and M. De Vos, "U-PASS: an Uncertainty-guided deep learning Pipeline for Automated Sleep Staging," *Computers in Biology and Medicine*, vol. 171, pp. 108205, 2024.

ABSTRACT | Estimating uncertainty plays a vital role in enhancing reliability of machine learning systems by identifying areas of high and low confidence and reducing the risk of errors. This chapter introduces U-PASS, a specialized human-centered machine learning pipeline tailored for clinical applications, which effectively communicates uncertainty to clinical experts and collaborates with them to improve predictions. U-PASS incorporates uncertainty estimation at every stage of the process, including data acquisition, training, and model deployment. Training is divided into a supervised pre-training step and a semi-supervised recording-wise fine-tuning step. We apply U-PASS to the challenging task of sleep staging and demonstrate that it systematically improves performance at every stage. By optimizing the training dataset, actively seeking feedback from domain experts for informative samples, and deferring the most uncertain samples to experts, U-PASS achieves an impressive expert-level accuracy of 85% on a challenging clinical dataset of elderly sleep apnea patients. This represents a significant improvement over the starting point at 75% accuracy. The largest improvement gain is due to the deferral of uncertain epochs to a sleep expert. U-PASS presents a promising AI approach to incorporating uncertainty estimation in machine learning pipelines, improving their reliability and unlocking their potential in clinical settings.

5.1 Introduction

With the increasing prevalence of machine learning in healthcare, ensuring the safety and reliability of these systems is crucial. Uncertainty estimates can play a key role in achieving this goal at all stages of the machine learning workflow: during data acquisition, training, and deployment. In data acquisition and in training, it is important to evaluate the usefulness of the information that is utilized to train the machine learning model. Feeding too much data in a machine learning model is sub-optimal both financially and in terms of model performance [115, 116]. Medical data is often expensive to acquire, requiring specialized equipment to measure each parameter, and expert knowledge to interpret the data. Moreover, feeding uninformative data to the model can deteriorate the performance because it adds noise without adding information [117]. The same principles apply to active learning [118, 119], a human-in-the-loop training strategy in which the model actively queries an expert to label extra data points which are most informative to improve the model. At deployment time, typical deep learning models always make a prediction, no matter how uncertain they are. Such systems are not usable in high-stakes medical applications: if the machine learning model doesn't indicate when a prediction is uncertain, a clinician cannot rely on it, as all outputs are equally likely to be incorrect [120]. Consequently, we propose that uncertainty estimation provides a principled approach to solving the aforementioned challenges in order to successfully integrate a machine learning workflow in the clinic. We demonstrate how an uncertainty-guided machine learning pipeline can help in a sleep staging task on PSG recordings.

Sleep staging is characterized by high levels of uncertainty, as evidenced by its relatively low inter-rater agreement [50, 57]. The uncertainty arises from multiple sources. First, sleep staging rules are defined on healthy subjects, and state-of-the-art sleep staging models are mostly validated on healthy populations. They typically perform poorly on diseased subjects [50, 121]. Disease influences brain activity and, consequently, PSG recordings, creating uncertainty in the data. Second, models trained on data acquired with one measurement setup or device may not generalize well to data obtained using a different measurement protocol, resulting in the well-known distribution shift problem [45, 81, 96, 122]. Hence, measurement factors can also contribute to uncertainty. Leveraging uncertainty estimation tools can help identify these uncertainties, which are likely to cause poor-quality predictions. As such, uncertainty estimates enhance model transparency and allow to actively query a sleep expert to label uncertain

samples. Hence, by integrating uncertainty estimation into sleep staging, our uncertainty-guided deep learning pipeline for sleep staging (U-PASS) improves both trust and performance of the model.

A variety of uncertainty estimation methods have been studied in the context of machine learning and deep learning [123, 124]. Hüllermeier et al. [124] stressed the importance of distinguishing between two types of uncertainty, aleatoric and epistemic, and discussed attempts to quantify these in machine learning. These reviews [123, 124], like most works on uncertainty, focused on estimating uncertainty at deployment time. A number of recent works have also investigated the concept of uncertainty within sleep staging. Guillot et al. [125] investigated machine learning algorithm performances in comparison to inter-rater agreement among human experts, and found that machine learning algorithms trained on multi-scored datasets can outperform clinical experts. Stephansen et al. [126] introduced hypnodensity graphs to report sleep stages along with their probabilities, as an alternative to the widely used hypnograms. These hypnodensity graphs were shown to correlate with the average scoring of expert panels [127, 128], and sleep staging models were shown to benefit from training on probabilistic, 'smoothed', labels based on these average scores [129]. Alternative methods to measure uncertainty in sleep stage scoring were based on the entropy of the output predictions [50, 130], the margin between the probability output of the most likely and second-most likely class [131], and Monte-Carlo dropout [132]. These works addressed the important issue of uncertainty quantification in sleep staging, but were largely focused on post-hoc uncertainty quantification at deployment time, and did not discriminate between different types of uncertainty. Going further than quantifying uncertainty, Kang et al. [130] explored whether sleep experts could successfully correct the labels of uncertain samples. A few studies also investigated uncertainty in the context of active learning in sleep staging [133]. MacAs et al. [131, 134] explored an uncertainty-based active learning approach for sleep staging, and Grimova et al. [135] proposed weighing uncertainty with density measures.

In contrast to previous works mentioned, this chapter integrates different types of uncertainty throughout the whole machine learning pipeline. The goal here is to take a leap forward, leveraging uncertainty estimates to enhance the performance and reliability of machine learning models and to effectively combine them with human capabilities. Our framework aims to improve predictions at every stage of the machine learning process. It should be model-agnostic, as it needs to be applicable to various clinical machine learning models and problems. We fulfill these goals by building an uncertainty-guided machine learning pipeline. Our pipeline follows a data-centric approach, and integrates uncertainty into data acquisition, training, active learning, and model deployment. We show how this greatly improves machine learning predictions and overall model reliability at

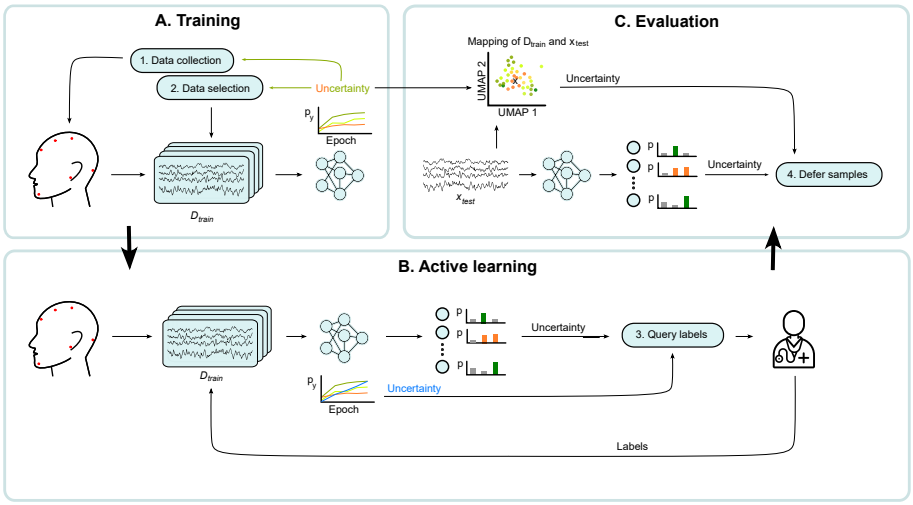


Figure 5.1: Full uncertainty-guided pipeline for sleep staging (U-PASS). The pipeline is more generally applicable to machine learning for clinical applications. **A.** At training time, data uncertainty is tracked to (1) select which channels (or features) to use, and (2) remove uninformative data. **B.** After training, active learning is used to fine-tune the model. Model uncertainty is tracked to (3) identify the most informative samples to query. **C.** At evaluation or deployment time, the most uncertain samples are deferred to the clinician (4). Used symbols: p is the model output, p_y is the model output for the true class y, D_{train} is the training dataset and x_{test} is a test sample.

different levels:

1. we improve the model through the quality of the training data,
2. we fine-tune the model to make better predictions on difficult examples using feedback from the clinician, and
3. we reduce mistakes by deferring uncertain predictions to the clinician.

Combining the capabilities of AI and clinicians, our pipeline represents a crucial step towards the integration of machine learning in clinical settings.

5.2 Methods

5.2.1 Data and study cohorts

This chapter uses all 90 PSG recordings of the Leuven - SDS dataset (see Section 2.1.2). In this chapter, we only use the C4-A1, F4-A1 and O2-A1 EEG derivations, as well as the chin EMG and the left-right EOG derivation, all resampled at 100 Hz. 45 recordings are used as a train set, and 45 recordings are used as the test set and for active learning.

5.2.2 Model and training

U-PASS can be applied to any machine learning model and problem, as we designed it to be model agnostic, with no requirements or adaptations to the network architecture or training procedure. In this paper, we use it in a sleep staging context with the SeqSleepNet model [46] (see Section 2.2).

5.2.3 Formulation of uncertainty

U-PASS utilizes insights from two recent methods that examine the characterization of uncertainty in training data. These methods analyze the training dynamics, which refers to the behavior of individual samples during the model training process. Data Maps [136] maps instances by tracking predictive confidence and model uncertainty (epistemic uncertainty) during training. DataIQ [137] on the other hand maps instances using predictive confidence and data uncertainty (aleatoric uncertainty). Both methods use mappings to stratify the training dataset into subgroups of easy, hard and ambiguous data. We further refer to instances with high epistemic uncertainty as *model-ambiguous* samples. These are examples for which the prediction variability of the model is highest during training. They are important to improve the performance and generalizability of a model [136]. On the other hand, instances with high aleatoric uncertainty are further referred to as *data-ambiguous* samples. These are samples for which predictions are the most uncertain on average, and on which the model systematically underperforms [137].

Let us formally define the characterization of model-ambiguous samples and data-ambiguous samples. Consider a typical supervised learning setting for a classification problem, where inputs $x \in \mathbb{R}^d$ need to be assigned to classes $y \in \mathbb{N}$. The goal is to learn a model f_θ which maps inputs to outputs by assigning a probability to each class given the input: $f_\theta(x) = P(Y|X = x, \Theta = \theta)$. Let

θ designate a certain instantiation of the model parameters. The model is iteratively trained for e^* training epochs, resulting in e^* different instantiations of the model parameters $\{\theta_1, \theta_2, ..., \theta_{e^*}\}$. Let $\Theta \sim P_{emp}(\{\theta_1, \theta_2, ..., \theta_{e^*}\})$ be a random variable with an empirical distribution over the model parameters throughout the training process. $\mathcal{P}(x, \theta) = [f_\theta(x)]_y$ is the model's output probability for the true class label. Using the variance as an uncertainty metric, we can decompose the uncertainty by the law of total variance. The epistemic uncertainty of the prediction is $v_{ep} = \mathbb{V}_\Theta[\mathcal{P}(x, \Theta)]$. The variance is evaluated over the model outputs during training, hence epistemic uncertainty captures the variability of model outputs, representing model uncertainty. The aleatoric uncertainty is $v_{al} = \mathbb{E}_\Theta[\mathcal{P}(x, \Theta)(1 - \mathcal{P}(x, \Theta))]$ [137]. In the aleatoric uncertainty, the variance is computed over the prediction. Hence, this represents the data uncertainty. Using the empirical distribution of Θ, the uncertainties are approximated as follows [137]:

$$v_{ep} = \frac{1}{e^*} \sum_{e=1}^{e^*} (\mathcal{P}(x, \theta_e) - \overline{\mathcal{P}}(x))^2, \tag{5.1}$$

$$v_{al} = \frac{1}{e^*} \sum_{e=1}^{e^*} \mathcal{P}(x, \theta_e)(1 - \mathcal{P}(x, \theta_e)), \tag{5.2}$$

with $\overline{\mathcal{P}}(x) = 1/e^* \sum_{e=1}^{e^*} \mathcal{P}(x, \theta_e)$.

We define model-ambiguous data as data with a high v_{ep}, and data-ambiguous samples as data with high v_{al}. In making these two groups, we use the definitions of 'ambiguity' proposed in Data Maps [136] and Data-IQ [137], respectively. We can further stratify the training dataset using the concept of predictive confidence, defined in both works as the mean output probability for the ground-truth-class:

$$c = \overline{\mathcal{P}}(x), \tag{5.3}$$

i.e. the confidence in the correct class. This allows us to define easy-to-classify samples as instances with high confidence and low ambiguity and hard-to-classify samples as instances with low confidence and low ambiguity.

5.2.4 U-PASS pipeline

Based on the uncertainty quantification methods from Section 5.2.3, we build a full uncertainty-guided pipeline, outlined in Figure 5.1. **A.** First, we tailor and improve the training dataset. Even before collecting a full dataset, we can

start training a model using various measured signals and monitor the data uncertainty (5.2) to pinpoint which signals are effective in decreasing uncertainty. This information can aid in deciding which modalities and channels should be recorded. If a particular signal does not lower the data uncertainty, it should be excluded. Furthermore, there is a trade-off to be made between the cost of data acquisition and the amount of uncertainty that can be tolerated to achieve a clinically useful result. After choosing a measurement setup, a full dataset is collected. We can then train a model on this dataset, and track the data uncertainty to identify the most data-ambiguous samples. These samples are removed from the training dataset, as they decrease the generalization performance of the model instead of improving it [137]. **B.** After training, the second step in U-PASS is to fine-tune the model using active learning. The model uncertainty can be used to identify for which recordings fine-tuning is the most useful, i.e. which recordings have the highest learning potential. Indeed, model uncertainty measures how much predictions of the model vary during training. In this chapter, we achieve fine-tuning with the semi-supervised adversarial domain adaptation from Chapter 4, using the additional labels supplied through active learning. **C.** Lastly, we can use uncertainty at evaluation or test time. A simple uncertainty measure based on uncertainty of the neighboring training samples can be used to identify uncertain test samples. When the probability of making an incorrect prediction is high, it is preferable to 'reject' a sample and defer it to a clinician rather than risk an incorrect prediction. However, deferring a prediction to a clinician incurs a certain cost. Therefore, there is a trade-off between the 'risk' or accuracy and the 'coverage' or percentage of samples deferred [138, 139].

The three principled uncertainty-guided steps in U-PASS result in a reliable and robust machine learning pipeline that is tailored for clinical environments. This framework enables effective interaction between clinicians and the model, thus enhancing clinical applicability. We delve into each step next.

Data collection and data selection

In the data collection experiments, we train the model on different numbers of channels to identify the best measurement setup, i.e. the measurement setup with the lowest data uncertainty. In the data selection experiments, we train on the best channels based on the data collection experiment, removing different amounts of ambiguous training data in each experiment. In both the data collection and data selection experiments, the model is trained on the full training dataset for 10 training epochs. Early stopping is applied using a validation dataset consisting of 2 recordings, retaining the model with the best performance on the validation set. As the training dataset consists of

45 recordings, each experiment is repeated $\lfloor 45/2 \rfloor = 23$ times, with different recordings in the validation set for every repeat. Accuracies are computed on the separate test set consisting of 45 recordings. The test set accuracies and training set uncertainty measures reported in Figure 5.3a are computed as averages over these 23 repeats.

Active learning

The active learning experiments start from the model obtained with the optimal setting based on the channel and data selection experiments. Using active learning to query informative labels, the model is then fine-tuned to personalize it to individual patients with the semi-supervised adversarial domain adaptation approach from Chapter 4. The training set of 45 patients is used as the source dataset, and each selected individual recording of the test set as the target dataset for adversarial domain adaptation. This experiment is performed 23 times, with each of the trained models. The labels for the semi-supervised training approach are acquired through the following active learning scheme. Before every training epoch, 1% of samples (roughly 10) are queried. The labels of these samples are used during that training epoch, along with the labels cumulatively acquired in previous epochs. The basic unsupervised adversarial domain adaptation is thus augmented by adding a supervised loss from the queried labels. The final model obtained after 10 training epochs with querying is retained for evaluation. As the same recordings are unavoidably used for training and testing in this active learning personalization step, evaluation is only performed on the unlabeled portion of those recordings.

The model uncertainty is a measure for how much the model performance varies while training on certain data, and can hence be seen as a measure for learning potential. Therefore, we focus the labeling efforts of the clinician on the recordings with the highest model uncertainty. Section 5.2.3 describes how to estimate model uncertainty in (5.1) using the variance of the prediction for the correct class as an uncertainty metric. This is applicable specifically for training data that has ground truth labels available. In our active learning scenario, labels are not available. However, we can still analyze training dynamics to estimate uncertainty, using entropy as an alternative, unsupervised uncertainty metric [140, 141]. The definitions of epistemic (model) and aleatoric (data) uncertainty with this uncertainty metric are the following:

$$v_{al}^{unsup} = \mathbb{E}_{\Theta}[H[P(y|x,\Theta)]] \tag{5.4}$$

$$v_{ep}^{unsup} = H[\mathbb{E}_{\Theta}[P(y|x,\Theta)]] - \mathbb{E}_{\Theta}[H[P(y|x,\Theta)]] \tag{5.5}$$

with H the entropy metric, and $P(y|x, \Theta)$ the output distribution. The expectations are again approximated with averages over the model parameters as in (5.1) and (5.2). We train the model on every recording using unsupervised adversarial domain adaptation, estimate the model uncertainty v_{ep}^{unsup} (A.2) based on the training dynamics in this unsupervised training step, and select the recordings with the highest v_{ep}^{unsup} (A.2) for active learning. The querying strategy to select the samples that are most informative is based on the same uncertainty estimate, but any other querying strategy could also fit in the U-PASS pipeline. We shall further refer to this approximation of model uncertainty (A.2) as the unsupervised model uncertainty, to make a distinction with the model uncertainty defined in (5.1).

Deferring ambiguous test samples

In the evaluation experiments, we use different models. For the selected recordings with the highest model uncertainty, we use the personalized models obtained in the active learning experiments. Evaluation is performed on the unlabeled portion of those recordings. For the other recordings, we use the optimal non-retrained models obtained from step A in Figure 5.1. As uncertainty metrics based on training dynamics cannot be used at deployment time, the evaluation experiments rely on post-hoc uncertainty metrics. Several post-hoc uncertainty metrics are compared, including two that are based on the output of the model: the entropy of the output distribution and the maximum output probability (i.e. the output probability of the predicted class). Two other metrics are distance-based. One metric measures the distance of the test sample to its n closest train samples, while the other uses the distance of the test sample to the n closest training samples for each separate class, and is calculated as the proportion of the smallest distance to the second-smallest distance. The remaining three metrics calculate uncertainty by averaging uncertainty metrics of the n closest training samples, weighted by their distance to those training samples. One metric uses the average confidence (5.3) of the closest training samples, the second uses the average data uncertainty (5.2), and the third uses the average model uncertainty (5.1).

Every uncertainty metric for evaluating uncertainty on the test set has its own scale, and it is up to the domain experts to define what level of uncertainty and how many mistakes can be tolerated. Therefore, this experiment ranks test set samples from most to least uncertain, and plots the accuracy corresponding with the z% most uncertain test samples. The uncertainty measures are then compared by averaging these accuracies over the 23 repeats.

The code for U-PASS is available in Github and can be accessed via this link: github.com/elisabethRMH/U-PASS.

5.3 Results

5.3.1 Insights from uncertainty

First, we illustrate how the two types of uncertainty are manifested in our data in Figure 5.2. Figure 5.2a shows the training dynamics for the 1% most model-ambiguous and data-ambiguous samples, as well as for the 1% most easy-to-classify and hard-to-classify samples. Model-ambiguous samples show the greatest learning curve during training, while data-ambiguous samples are characterized by a flat learning curve. The distributions of the two types of ambiguity are visualized in Figure 5.2b. The figure shows two-dimensional UMAP embeddings of the training data, with the different colors indicating the five sleep stages in the left figure, and heatmaps showing the model and data uncertainty in the middle and right figure, respectively. In both heatmaps, the 1% most ambiguous samples are highlighted in black. The 1% most model-ambiguous samples are found mostly at the borders and outside the main training data distribution, as well as on class boundaries. Indeed, atypical and out-of-distribution samples are more challenging for the model at first, but the model can learn these samples through more training. The most data-ambiguous samples are centered in sleep stage N1, which is the least well defined and hardest to classify. Data uncertainty is also higher on the class boundaries, where some samples contain characteristics of multiple sleep stages.

5.3.2 Data collection

U-PASS starts with evaluating the data uncertainty during data collection to assess which signals should be acquired and used. Figure 5.3a shows the confidence and data uncertainty in the PSG training datasets, mapped by using Data-IQ [137]. Each 30-second sample is one dot, and the density of samples is represented by the brightness of colors. The figure demonstrates that most samples are in the upper left corner, characterized by high confidence and low data uncertainty (easy-to-classify). Some samples have high data uncertainty (data-ambiguous), and a minority of samples is characterized by low data uncertainty and low confidence (hard-to-classify). The concentration of samples in the data-ambiguous region decreases when going from one channel to three channels, and decreases again from three channels to five channels. This is

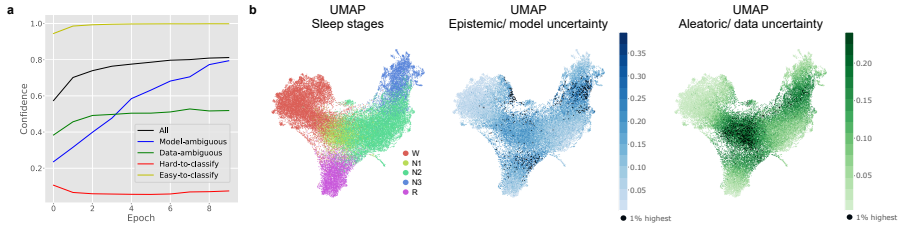


Figure 5.2: (a) The average training dynamics for the 1% most data-ambiguous, model-ambiguous, easy-to-classify and hard-to-classify samples. (b) UMAP visualization of the feature distribution resulting from training a sleep staging network. The left plot shows the sleep stages in this distribution. In the middle plot, the tint indicates the model uncertainty, and the 1% most model-ambiguous are shown in black. In the right plot, the tint indicates the data uncertainty, and the 1% most data-ambiguous are shown in black.

also evident from the average data uncertainty over the whole training dataset, which is shown under the plots. The average confidence increases accordingly, as samples go from being data-ambiguous to easy-to-classify. As a result, the accuracy on the test set also improves, which is the final proof that reducing the data uncertainty is helping the outcome.

We conclude that tracking the data uncertainty allows to compare the quality of different measurement setups for sleep monitoring. Indeed, a decrease in data uncertainty with the addition of a channel indicates that this channel adds valuable information. It can advise users in selecting the optimal setup that will result in the best accuracy at deployment time.

5.3.3 Data selection

We can curate our training dataset, by discarding from it the most data-ambiguous training samples. Figure 5.3a shows how removing these ambiguous data affects the model's test performance on an independent test set. There is a trade-off to be made between removing samples that confuse the model and removing too many samples so that useful information is discarded from the training dataset. In this case, the optimal point is reached when discarding 1% of training samples.

Data uncertainty in sleep staging can arise from biological sources (e.g. age, pathology) and sources related to the measurement technique (e.g. segmentation, interference) [57], as well as from ambiguity in sleep scoring rules. Figure 5.4a shows that data uncertainty per segment is significantly higher for segments on

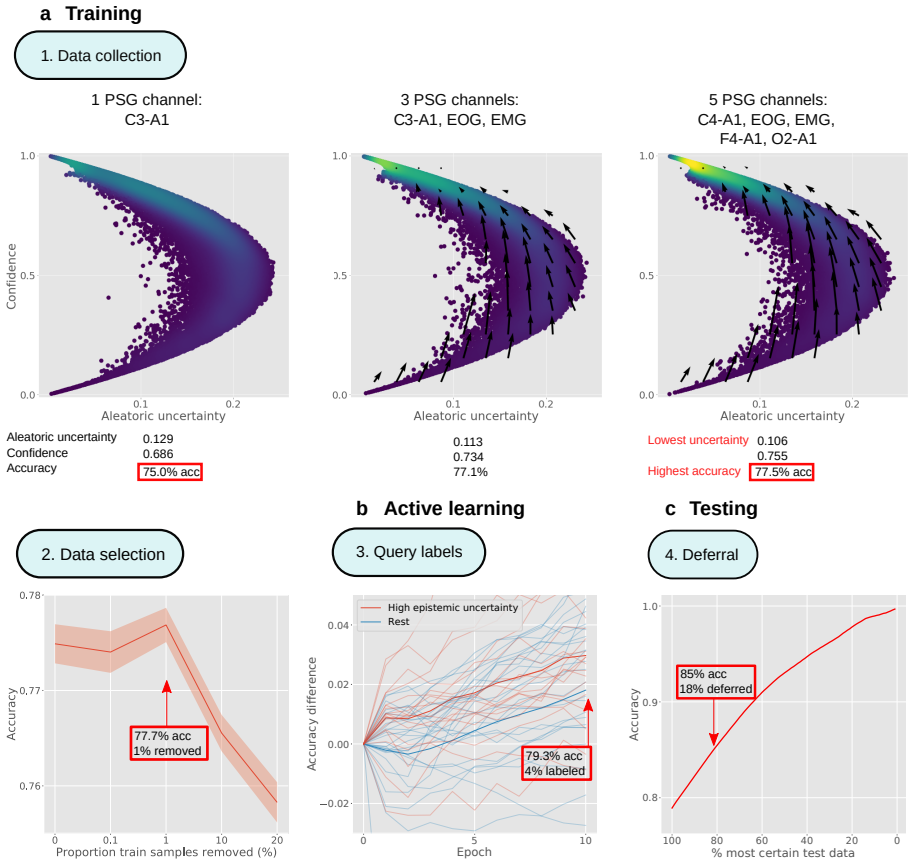


Figure 5.3: U-PASS applied to hospital-based PSG data. The resulting accuracy after every consecutive step is indicated with a red box. (a) During training, the training dataset is tailored using data collection and data selection based on data uncertainty. In the data collection step, samples move to the upper-left corner with high confidence and low data uncertainty when more channels are used. We thus proceed with all five channels. In the data selection step, the optimal test result is achieved when removing the one percent (roughly 440) most data-ambiguous training samples. (b) Active learning is used to adapt the model to individual recordings. 1% of samples is queried in each of the 10 training epochs. We only perform active learning on recordings characterized by high model uncertainty. On the rest of the recordings, it doesn't show a large improvement. (c) At deployment time, uncertainty-based deferral increases the test accuracy to the desired level of 85%. Samples with uncertainty values below the threshold are referred to a clinician for manual labeling.

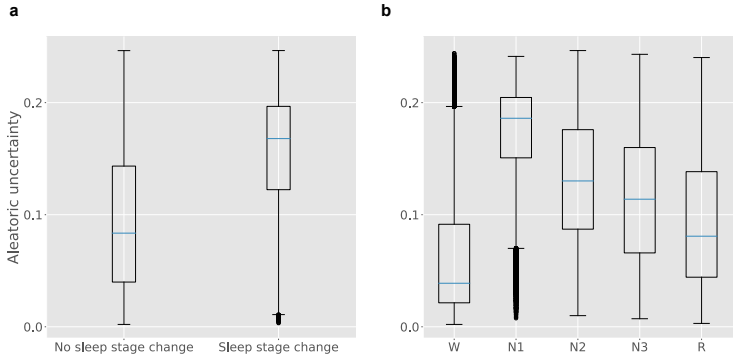


Figure 5.4: Parameters causing data uncertainty in the PSG training dataset. (a) The data uncertainty per segment is higher for segments on sleep stage transitions than for segments that are not on sleep stage transitions. (b) The data uncertainty per segment, aggregated per sleep stage, shows that wake segments have low data uncertainty and N1 segments have the highest data uncertainty.

sleep stage transitions than for segments that are not on sleep stage transitions. This source of uncertainty is hence related to the discretization caused by segmentation and the continuous nature of sleep, with features from multiple sleep stages in a single segment. Figure 5.4b shows that wake segments have the lowest data uncertainty, and N1 segments have the highest data uncertainty. The fact that different stages of sleep are characterized by different levels of uncertainty is related to how clearly sleep stages are recognized and defined, which is influenced by sleep scoring rules and the biology of sleep. We find evidence for this in the fact that the inter-rater variability between human scorers shows the same trends as data uncertainty [59]. This is further discussed in Section 5.4 and shown more formally with an experiment on a multi-scored dataset in Appendix 5.C.

5.3.4 Active learning

Once the model is trained on the improved training set, U-PASS evaluates unsupervised model uncertainty to assess on which recordings the model should be fine-tuned through active learning. We use active learning to personalize sleep staging models to those individual recordings. Figure 5.3b shows the average improvement from active learning to individual PSG recordings. The improvement is shown for the selected 40% recordings with the highest unsupervised model uncertainty and for the rest of the recordings. Unsupervised model uncertainty clearly points to the recordings that benefit the most from

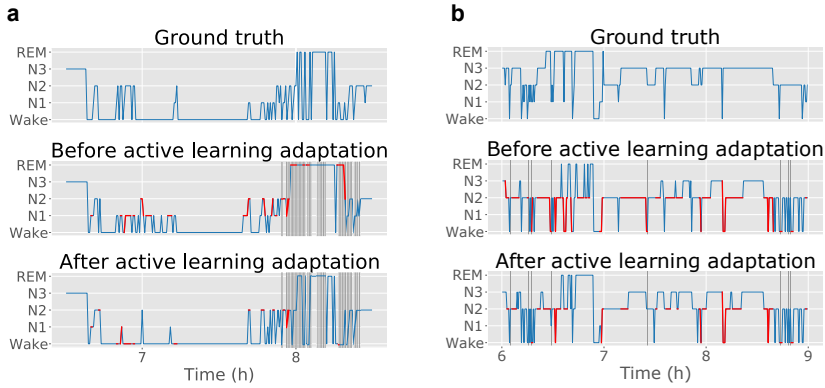


Figure 5.5: Hypnograms (sleep stages over time) for snippets of two different recordings, demonstrating how personalization through active learning (AL) improves the performance. Queried samples are shown in grey, labeling errors of the sleep staging model are shown in red. (a) Before AL, the sleep staging model incorrectly labels a lot of segments as N1 sleep around hour 7, and the jumps between REM-sleep and Wake around hour 8 are missed. After AL, most of these labeling errors are corrected. Although the queried samples are centered around hour 8, the mistakes in N1-sleep around hour 7 are partly corrected too. (b) Both the episode of REM-sleep between hour 6.5 and 7, and the episodes of N3-sleep between hour 7 and 8.5 are more accurately predicted after AL. Overall, we see that both the queried samples and the unlabeled parts of the recording are labeled more accurately after AL in both examples.

the active learning adaptation. Figure 5.5 shows two examples of how active learning improves the predictions in both labeled and unlabeled samples.

5.3.5 Deferring uncertain test samples

Lastly, the models are deployed on unseen data. For the 40% selected recordings with highest unsupervised model uncertainty, we use the personalized models. The other recordings are evaluated with the optimal non-retrained models from the data collection and data selection step. U-PASS then uses uncertainty measures to defer test samples to an expert when predictions are uncertain. Figure 5.3c shows the accuracy on the 100% to 1% most certain test samples. This is equivalent to the accuracy when removing 0% to 99% of the most uncertain test samples. The higher this accuracy, the better the uncertainty estimation, as the uncertainty metric should retain the most accurate predictions and defer the least accurate ones. Expressing this in terms of risk and coverage,

a curve that cuts a higher line achieves a more optimal risk-coverage trade-off [138, 139]. Different post-hoc uncertainty metrics are tested. Figure 5.3c only shows the best-performing one: the weighted average confidence of the n closest training samples. The performances of all the uncertainty metrics are shown in Supplementary Figure 5.A.

The results in Figure 5.A show that different uncertainty metrics perform best depending on the threshold set on the tolerated uncertainty. If we require 85% accuracy, the confidence of neighboring training samples achieves the best performance, with 18% of samples deferred. For reference, 85% accuracy is a great performance, as the inter-rater agreement in sleep staging is estimated to be 82.6% [58, 59]. When we exclude samples that lie on sleep stage transitions, the accuracy improves from 85% to 89%.

Selecting an operating point on the risk-coverage curve (Figure 5.3c) involves balancing the trade-off between the cost of errors and the cost of human labeling. The 85% accuracy operating point is a safe choice, as it is slightly higher than the inter-rater agreement among experts for this clinical application. However, from a health-economic perspective, knowing the cost of human labeling would allow us to make a more informed decision and potentially select a more optimal operating point.

As clear from Figure 5.A, the difference between the entropy metric and metrics based on the uncertainty in the training set is mostly quite small, and depends on the uncertainty threshold we set. However, the metrics based on uncertainty in the training set have an important advantage compared to the decision entropy, as they provide a reason for deferral. Indeed, the distance-weighted confidence of neighboring training samples is a product of the distance from the training samples and the uncertainty of close training samples. Figure 5.6 visualizes both factors. The Spearman correlation coefficient shows that accuracy is negatively correlated with distance from the closest training samples ($r(43)=-.40$, $p=.0067$) and positively correlated with the confidence of the closest training samples ($r(43)=.51$, $p=.00032$). The two factors detect, respectively, out-of-distribution samples, and samples lying close to uncertain regions, e.g. decision boundaries.

The effect of applying U-PASS for sleep staging is summarized in the red boxes of Figure 5.3, which show the resulting accuracy after each step in the pipeline. This clearly demonstrates how U-PASS leverages uncertainty in training, active learning, and deployment to improve the accuracy at every stage. Appendix 5.B includes an additional experiment on the Dreem Open Datasets [125], showing that our pipeline also generalizes to other datasets.

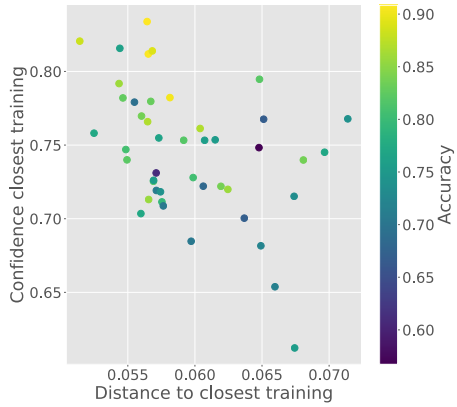


Figure 5.6: The accuracy on the test data depends on both the distance from the training data and the uncertainty of those training data. For every patient, this figure shows the accuracy, confidence of the closest training samples and the distance to those training samples. These metrics are averaged over all the samples of the patient's recording, so every patient is represented by one dot.

5.4 Discussion

We have developed and validated U-PASS, an uncertainty-guided pipeline for automated sleep staging. Although we applied U-PASS to sleep staging, the pipeline is generally applicable to any machine learning problem. It is particularly useful for machine learning in healthcare, which has strong requirements for reliability, trust, and benefits from user interaction. We have demonstrated that U-PASS can incrementally improve predictions by utilizing data uncertainty and model uncertainty at different stages of the data processing pipeline (training, active learning, and deployment) and by interacting with clinicians. It can easily be tailored to the accuracy needs of any particular application, by adapting the amount of data queried and deferred in the active learning and deployment phase, respectively.

We have validated our pipeline in PSG data of a real patient population of elderly subjects with suspicion of sleep apnea. Many performant deep learning models have been developed for sleep staging on PSG, but they are developed as static, stand-alone input-output machines. This limits their performance and utility in the clinic. Our contribution consists of building a pipeline in which such deep learning models can be integrated, boosting their performance and allowing for the clinician to choose the desired performance level. To showcase the broad applicability of our approach and its resilience in scenarios

involving domain shifts, we conducted additional validation of our pipeline on a multi-scored, openly accessible sleep dataset (Appendix 5.B and 5.C).

The performance gains achieved by applying U-PASS to a state-of-the-art sleep staging model are shown in Figure 5.3. The first two parts of the pipeline, data collection and data selection, are focused on tailoring the training dataset to achieve optimal results. Both steps decrease the data uncertainty in different ways. In the data collection step, data uncertainty is estimated to find out which features (or channels in the case of sleep staging) need to be collected. Once the data acquisition protocol is fixed and the training set is acquired, the data selection step filters out the data with the highest data uncertainty, as these can deteriorate model performance. In our experiments, the data collection step shows the largest improvement by going from one EEG channel to five channels. The data selection attains its maximum improvement when 1% of the training data are discarded. This step only increases the accuracy by a modest relative 0.3%. Despite the minor magnitude of the improvement, we systematically observe improvements with data pruning in other datasets. Appendix 5.B shows the result on a public dataset. The current data selection approach is still rather rudimentary and may benefit from further refinement. First, it does not account for class imbalance. In the case of sleep staging, this particularly concerns class N1, the smallest class with samples often found on class boundaries. Caution is advised to avoid compromising N1 sensitivity and future work could explore balancing the removed samples across classes. Second, in the context of sleep staging, the current data selection approach does not consider temporal dependencies between consecutive epochs. It remains an open question how to optimize data selection for time series with temporal dependencies. Future research could investigate how to take these into account. Future investigations could explore the impact of various factors, such as data type, quality, and volume, as well as the model architecture, on the effectiveness of data selection.

The third part of the U-PASS pipeline includes fine-tuning a sleep staging model with an active learning approach. Active learning or any other form of additional training can resolve model uncertainty, which is caused by distribution shifts (related to patient populations or measurement equipment) or out-of-distribution data. Another and perhaps even more important source of model uncertainty is the difference in interpretations of scoring rules between different experts or sleep centers [142]. We have developed a recording-wise personalization step to handle out-of-distribution data, or 'difficult-to-score' recordings and samples. Adaptation to a new scorer is not the focus here, as our experiments are performed on a dataset labeled by a single scorer, but this could also be resolved through additional learning. Variability due to hardware and different patient populations is addressed in Appendix 5.B, which shows that the pipeline

also handles domain shifts between the training and test set.

The last step of U-PASS, deferring ambiguous samples, allows us to choose a tolerated level of uncertainty or a desired accuracy for a specific task. This comes at a labeling cost, similar to the active learning step of the U-PASS pipeline. The difference between the two steps is the type of uncertainty they tackle. Active learning can only resolve model uncertainty, coming from a lack of knowledge of the model. Pure data uncertainty, on the other hand, is inherent to the data and can hence not be resolved through learning. The only possible mechanism to cope with high data uncertainty at deployment time is deferral to an expert. In theory, doctors do not know more than a machine learning model in cases of pure data uncertainty. However, they can take the right course of action, whether that be ignoring the data, performing a new measurement, or fixing a broken electrode. We conclude that both active learning and deferring help to increase the accuracy to the desired level at a certain labeling cost. Since they tackle different types of uncertainty, they should be combined for optimal performance and utility.

Along with performance gains, the uncertainty estimation methods in the U-PASS pipeline also provide some interesting clinical insights into the sleep data. Figure 5.2b and Figure 5.4 shows that the data uncertainty in segments on sleep stage transitions is higher than for others, and that sleep stage N1 is characterized by more uncertainty than other sleep stages. This data uncertainty is by definition inherent to the data and can not be resolved by more training. Hence, human expert annotations should exhibit the same data uncertainty, and indeed, our results are consistent with the sources of uncertainty identified in inter-rater variability studies [59]. In line with our results, previous studies on automated sleep staging found that hypnodensity plots, showing the model's output probabilities for all sleep stages, also correlated well with the average scoring of expert panels [126–128].

Sleep scoring rules are notoriously ambiguous, leading to high variability between scorers and sleep centers [121, 142]. As such, the objectivity of uncertainty metrics can guide clinical practitioners and experts developing sleep scoring guidelines [98] to define better rules. For example, changing the 30-second segmentation to a more fine-grained segmentation should benefit the data uncertainty. Furthermore, the less ambiguous the sleep staging rules are, the less data uncertainty there is. An additional insight we have gained from plotting the sleep features (Figure 5.2b) is that sleep stages don't seem to be discretely separated, but rather lie on a continuous spectrum. Recent works have advocated modeling sleep as a continuous process [143–145], which may be a more biologically accurate representation. Hence, uncertainty estimation methods can inform the medical field on how to best describe and define sleep.

5.5 Conclusion

This chapter presents U-PASS, a machine learning pipeline that integrates uncertainty-informed curation of the training set, uncertainty-based active learning to incorporate a clinician's feedback and deferral of uncertain decisions to a clinician. As such, all the facets of uncertainty and benefits of uncertainty estimation are harmonized in a single framework and combined with the clinician's capabilities. This optimizes the machine learning pipeline and unlocks its potential in a clinical setting by adding safety guardrails to the process. Not only does U-PASS give agency to clinicians, but it also has the potential to provide them with valuable insights into their data, offering a deeper understanding of sleep biology and potentially other clinical applications.

We have demonstrated U-PASS for sleep staging on PSG records. We envision its potential for use in machine learning-assisted sleep scoring within the daily clinical practice of sleep medicine, as well as in other clinical annotation tasks. Overall, U-PASS stands out as a promising approach for enhancing the reliability and safety of machine learning systems in critical healthcare settings.

Appendices

5.A Extended deferral results

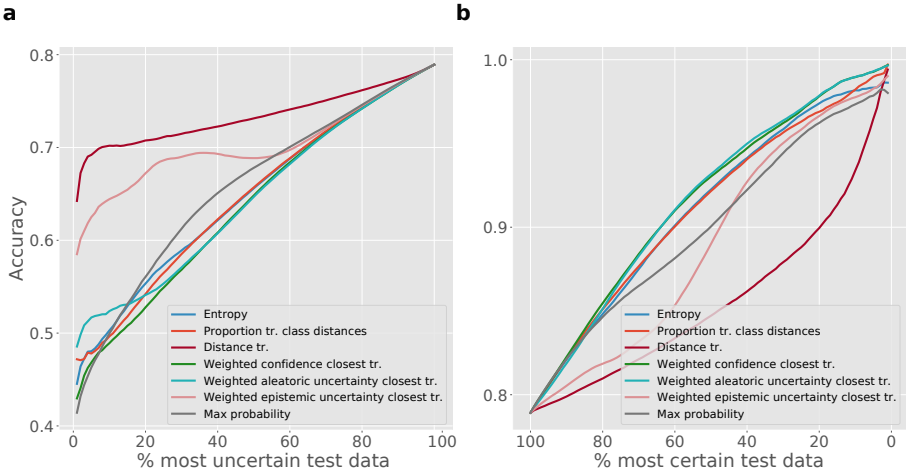


Figure 5.A: Extended results showing the performance of the different measures of uncertainty for deferring uncertain test samples. (a) The accuracy for the x% most uncertain samples. The better the uncertainty estimate, the lower this figure. (b) The accuracy for the x% most certain samples. The better the uncertainty estimate, the higher this figure. Depending on the threshold and scenario, different uncertainty measures perform best.

5.B U-PASS in Dreem Open Datasets

The U-PASS pipeline is applied to the Dreem Open Datasets (DOD) [125] to demonstrate its generalizability in different datasets. The chosen model architecture for sleep staging is still SeqSleepNet, and the consensus scores of this multi-scored dataset are used to train the model. It should be noted that a potential improvement here could be to train the model on smoothed labels, using probabilities obtained by averaging the scores over the expert panel [129]. The results for this experiment are shown in Figure 5.B. For the data collection and data selection experiments (Figure 5.Ba), the network architecture is trained on 28 randomly selected patients from DOD-O, the OSA cohort of the DOD. For the active learning personalization experiments (Figure 5.Bb) and evaluation experiments (Figure 5.Bc), we use the remaining 28 OSA patients from DOD-O and the 25 healthy participants from DOD-H. The DOD-H is a dataset with a

different patient population, acquired with different equipment, at a different institution compared to the DOD-O dataset. This dataset is included for testing to verify that the framework still works in case there is a domain shift between the training set and test set.

The results on the DOD are in line with the results on the SDS dataset (Figure 5.3), with a few interesting differences. In the data collection experiment (Figure 5.Ba, left), we only test a maximum of three channels, because these are the only ones that are present in both DOD-O and DOD-H. In the active learning experiments (Figure 5.Bb), the recording-wise personalization improves the accuracy more for the DOD-H patients than for the DOD-O patients. This is in line with our expectations, as active learning and other forms of training resolve model uncertainty, which is expected to be higher in DOD-H. Differences in measurement equipment and in patient characteristics typically lead to domain shift, and, consequently, model uncertainty and lower accuracies. The average unsupervised model uncertainty proves that epistemic uncertainty is indeed larger in the DOD-H dataset. At test time (Figure 5.Bc), we decide not to defer any samples from the DOD-O dataset, because the accuracy is near-perfect (greater than 85%) after the active learning step. For the DOD-H dataset, we defer 7% of samples to get to 85% accuracy. We also test what would happen if we choose an uncertainty threshold on the DOD-O data and apply it to the DOD-H data. When selecting an uncertainty threshold that results in 90% accuracy (with a deferral of 13% of samples) for the DOD-O test data, this threshold results in a deferral of 15% of samples for the DOD-H dataset, with a resulting accuracy of 87.5% on the remaining data. Although the uncertainty threshold set in one dataset can not guarantee the same accuracy in another dataset in case of domain shift, it still achieves a very decent result, with a final accuracy close to the one obtained in the original dataset. This shows that our pipeline is indeed robust to domain shift.

5.C Inter-rater variability

In general, inter-rater disagreement results from a mix of aleatoric (data) and epistemic (model) uncertainty. A study by Van Gorp et al. [57] showed that both machine learning models and humans suffer from both aleatoric and epistemic uncertainty. However, assuming that all the raters are well-trained, certified scorers, the contribution of aleatoric uncertainty is higher than the contribution of epistemic uncertainty. When there is an ensemble of predictions, we can take the expectation (average) over the different predictions to estimate aleatoric uncertainty, and the variance over the predictions to estimate epistemic uncertainty. This could in theory be done with a panel of scorers as the ‘ensemble’, but unfortunately, human scorers do not give

probabilities but one-hot predictions, which do not allow to calculate these estimates. Therefore, we can only get a measure of total uncertainty (aleatoric and epistemic) by taking the entropy or variance of the average prediction.

The Dreem Open Datasets were scored by 5 different scorers, allowing to quantify the total human uncertainty. Our training procedure for this dataset is explained in Appendix 5.B. For the 28 training recordings, we compared our estimates of uncertainty with the inter-rater variability. The inter-rater variability was calculated as $v_{tot} = P^* \times (1 - P^*)$ where P^* is average rating of the majority voted sleep stage. We obtained per-recording averages of inter-rater variability and uncertainty estimates, and calculated Pearson's correlation coefficient between these values. The best match was found between inter-rater variability and the data uncertainty (5.2) estimated on our training data, with a correlation coefficient of r=0.64 (see Figure 5.C). The correlation coefficient obtained with model uncertainty (5.1) was r=0.56. Model uncertainty is model-dependent, and so, it is not necessarily the same in machine learning models and humans. In conclusion, although this experiment does not exclude the presence of epistemic uncertainty in the human scorers, it does confirm that data uncertainty is model-independent and thus correlates very well between human scorers and machine learning models.

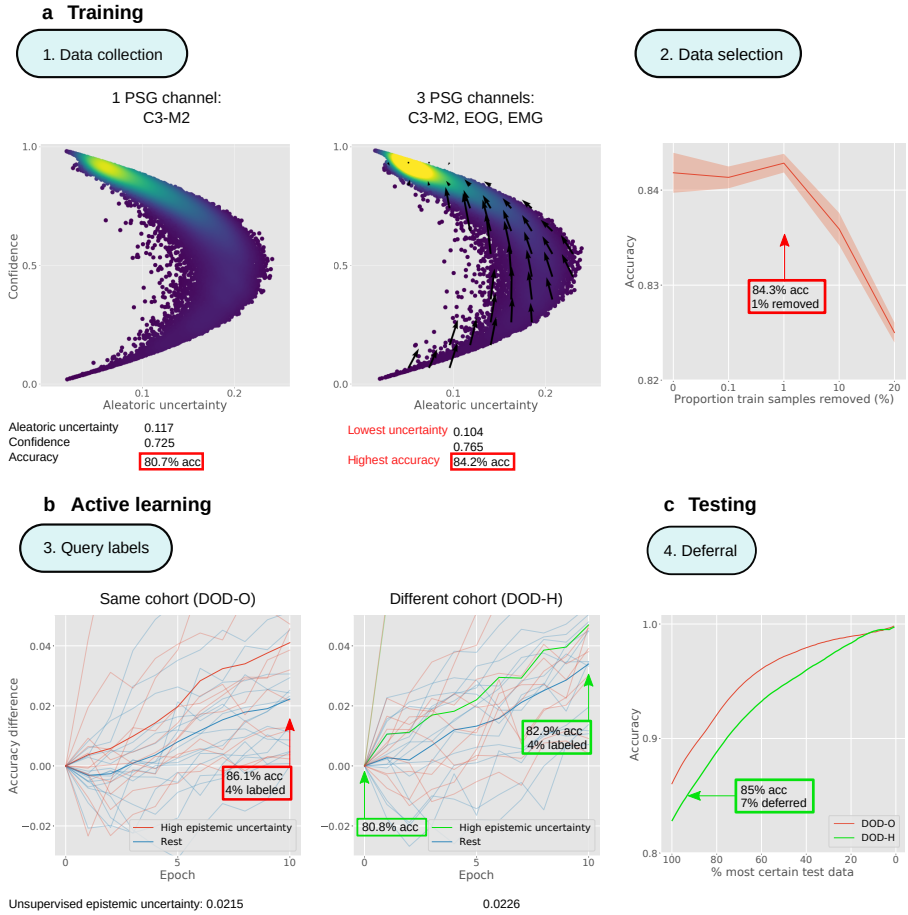


Figure 5.B: U-PASS applied to the Dreem Open Datasets (DOD). The resulting accuracy after every consecutive step is indicated with a box. (a) During training, the training dataset is tailored based on data uncertainty. In the data collection step, samples move to the upper-left corner with high confidence and low data uncertainty when more channels are used. We thus proceed with three channels. In the data selection step, the optimal test result is achieved when removing the one percent most data-ambiguous training samples. (b) Active learning is used to adapt the model to individual recordings. 1% of samples is queried in each of the 10 training epochs. Active learning shows a larger improvement on DOD-H than on DOD-O, which is as expected because of the domain mismatch. In both datasets, we only perform active learning on the 40% recordings characterized by high model uncertainty. (c) At deployment time, uncertainty-based deferral increases the test accuracy in DOD-H to the desired level of 85%. Samples with uncertainty values below the threshold are referred to a clinician for manual labeling. In the DOD-O dataset, this accuracy is already achieved without deferral, so we do not defer any samples.

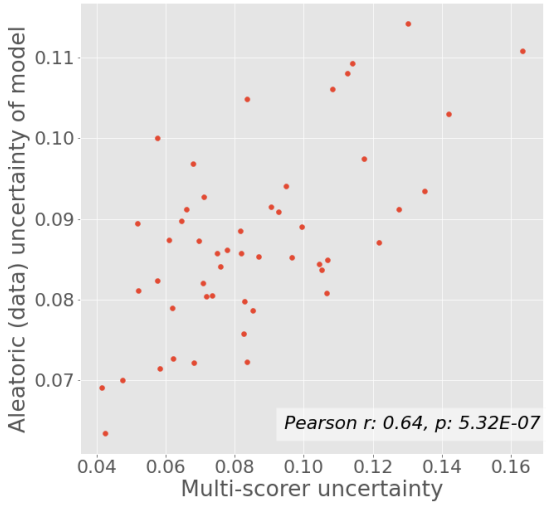


Figure 5.C: The relation between the data uncertainty ('aleatoric uncertainty') and the inter-scorer variability for 28 randomly selected recordings used for training U-PASS on the DOD-O dataset.

6 | Automated remote sleep monitoring needs uncertainty quantification

This chapter is largely based on **E. R. M. Heremans**, L. Van den Bulcke, N. Seedat, A. Devulder, P. Borzée, B. Buyse, D. Testelmans, M. Van Den Bossche, M. van der Schaar, and M. De Vos, "Automated remote sleep monitoring needs uncertainty quantification," *Journal of Sleep Research*, e14300, 2024.

ABSTRACT | In this chapter, we used U-PASS, the uncertainty-centered machine learning pipeline from Chapter 5, to automate sleep staging in a challenging real-world dataset of single-channel EEG and ACM collected with a wearable device from an elderly population. We were able to effectively limit the uncertainty of our machine learning model and to reliably inform clinical experts of which predictions were uncertain to improve the machine learning model's reliability. This increased the five-stage sleep scoring accuracy of our sleep staging model from 63.9% to 71.2%. Remarkably, the machine learning approach outperformed the human expert in interpreting these wearable data. Manual review by sleep specialists, without specific training for sleep staging on wearable EEG, proved ineffective. The clinical utility of our automated pipeline was also demonstrated, establishing a strong correlation between the predicted sleep parameters and the reference PSG parameters, and reproducing known correlations with the apnea-hypopnea index. In essence, this chapter presents a promising avenue to revolutionize remote patient care through the power of machine learning enhanced with uncertainty estimation.

6.1 Introduction

Chapter 5 discussed sources of uncertainty and highlighted the importance of uncertainty quantification in PSG-based sleep staging. However, uncertainty quantification in wearable-based sleep staging also warrants attention, as the challenges are different – and in some aspects exacerbated – compared to those in PSG. The same biological factors that introduce uncertainty in PSG are present in wearable recordings. Additionally, measurement-related uncertainties are expected to be greater due to generally lower-quality recordings. As wearables are often self-applied and more comfortable, they can suffer from issues like incorrect device application or detached electrodes, leading to unusable signals. Furthermore, with the limited amount of labeled data available, epistemic uncertainty (because of incomplete model knowledge) may be higher. Lastly, compared to PSG, wearable-based sleep staging might introduce new challenges when clinicians evaluate uncertain epochs, since they struggle to interpret signals acquired from non-standard EEG positions [35].

The present chapter investigates how U-PASS can facilitate remote sleep monitoring with wearable devices (Figure 6.1). Specifically, it focuses on the following objectives:

1. Improve sleep staging accuracy in a challenging real-world dataset, consisting of single-channel wearable EEG and ACM collected from an elderly population with suspected sleep apnea.
2. Enhance the machine learning pipeline by excluding low-quality training data during model training and flagging uncertain data during deployment. These steps, respectively, are aimed at improving the model and reducing mistakes.
3. Investigate the role of human annotators in remote sleep monitoring with wearables, as compared to in-hospital standardized PSG studies.
4. Verify that a machine learning model based on behind-the-ear wearable EEG picks up on relevant diagnostic features with enough accuracy to be clinically useful.

By addressing these aims, this chapter seeks to demonstrate the potential of integrating wearable technology into sleep monitoring practices, providing insights that could facilitate more reliable and accessible solutions for sleep staging in clinical settings.

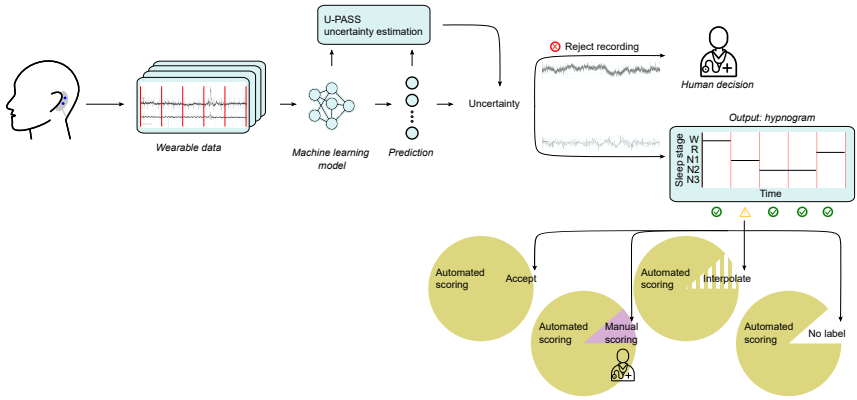


Figure 6.1: The proposed machine learning pipeline for automated sleep staging on wearable sleep recordings, enhanced with uncertainty quantification. Brain signals and potentially other physiological signals are measured from patients using a wearable device. After acquiring a dataset, a model is trained for sleep staging on these data. At training time, data uncertainty is tracked to select which channels to use, and to remove uninformative data. At evaluation time, the recordings with the most uncertainty are marked and presented to a clinician, who then decides whether to discard the recording or take further action. All other recordings are labeled by the machine learning model, and only the most uncertain samples of these recordings are marked. This chapter investigates what action is required for these uncertain samples: the four yellow circles show the four options we discuss.

6.2 Methods

6.2.1 Dataset and sleep staging model

In this chapter, we used the behind-the-ear sleep recordings from the Leuven - SDS dataset. As explained in Section 2.1.2, the wearable measurement of cross-head behind-the-ear EEG was simulated with standard EEG equipment in the first 45 recordings, and measured with the SensorDot in the last 45 recordings. The wearable device measurements also included angular velocity, and acceleration in all the recordings. The first 45 recordings were used as a training set, and the final 45 recordings were used as the test set. We trained SeqSleepNet [46] (see Section 2.2) to perform automated sleep staging on the wearable recordings in this dataset.

Table 6.1: Sample characteristics.

	Train dataset	Test dataset
n	45	45
Age	65.96 ± 4.95	66.13 ± 5.02
BMI*	33.44 ± 14.84	30.60 ± 6.50
% female	35.6	42.2
AHI	33.38 ± 21.20	35.17 ± 25.47

Abbreviations: BMI, body mass index. AHI, apnea-hypopnea index.
*BMI was not recorded for two patients of the test dataset.
Data are reported as mean ± standard deviation unless not applicable.

6.2.2 U-PASS and uncertainty measures

The sleep staging model was trained on the wearable data using the U-PASS uncertainty-guided pipeline from Chapter 5 (see Figure 6.1). We previously developed this pipeline to improve the training data and hence the model performance, and identify uncertainty in the test data at deployment time to increase reliability. At training time, U-PASS first determines the minimal measurements needed based on the amount of uncertainty. Second, it removes the most ambiguous training data to further improve the model performance. At test time, it flags uncertain data, increasing the model's reliability on the remaining predictions. In this chapter, this flagging at deployment time was performed at the level of whole recordings and at the sample level. First, particularly difficult recordings were marked and presented to the clinician for further decision-making. This step was a coarse preselection to discard recordings that were unreadable. Second, for the recordings that were considered reliable, the same uncertainty metric was deployed to flag individual segments for which the model prediction was uncertain.

To identify uncertainty in the training data and test data, U-PASS estimates uncertainty using predefined uncertainty measures. We used the measures defined in Chapter 5, which represented dynamics computed over training, rather than just capturing a single snapshot at the end of training. To identify uncertainty in the training data, we used the data uncertainty defined as $v_{al} = \mathbb{E}_{\Theta}[\mathcal{P}(x,\Theta)(1-\mathcal{P}(x,\Theta))]$ [137], where $\mathcal{P}(x,\theta)$ is the model's output probability for the true class, and Θ are the model's parameters over the model's training iterations. To identify uncertainty in the test data, we used the distance-weighted confidence of the neighboring training samples. For the n nearest training samples, the average confidence of the model is $c_i = \mathbb{E}_{\Theta}[\mathcal{P}(x_i,\Theta)]$ with $i = \{1...n\}$. For each of these training samples, the associated confidence was

weighted with its distance to the given test sample:

$$\sum_{i=1}^{n} \sigma(-d_i)\,(1-c_i)\,e^{\alpha\,\frac{1}{n}\sum_{i=1}^{n} d_i}, \quad \text{with } \sigma(-d_i) = \frac{e^{-d_i}}{\sum_{i=1}^{n} e^{-d_i}}. \tag{6.1}$$

$(1-c_i)$ is a measure of uncertainty of a training sample, with softmax weighting applied to give greater influence to the nearest samples. The exponential function $e^{\alpha\frac{1}{n}\sum_{i=1}^{n} d_i}$ incorporates the overall distance to all n training samples in the uncertainty metric. The two factors respectively identify test samples situated near uncertain regions, such as decision boundaries, and out-of-distribution samples[1]. Samples near decision boundaries typically exhibit features indicative of different sleep stages, for example, because of a sleep stage transition within an epoch. Out-of-distribution samples are samples that deviate from the training data, such as samples corrupted by a new type of noise. In general, the uncertainty in sleep staging emerges from biological factors such as disease and age, measurement technique-related factors like segmentation and interference [57], and different possible interpretations of sleep scoring rules. These factors are inherent problems to sleep staging, leading to significant inter-rater variability as well as uncertainty in machine learning models [59] (see also Chapter 5). Appendix 6.B visualizes some of these sources of uncertainty.

6.2.3 Manual review

The U-PASS pipeline allows to flag uncertain predictions at deployment time, but we have not yet determined what to do with those flagged instances. Figure 6.1 visualizes how uncertain instances are detected, and which different actions could be taken for these uncertain predictions. When applying machine learning on PSG in hospital-based sleep studies, uncertain decisions can be deferred to human certified sleep scorers who are trained to review PSG data [130] (see also Chapter 5). However, in the case of monitoring with wearables, the non-standard acquisition setups make the signals hard to read for clinicians [35]. In this chapter, we performed a small-scale experiment of preliminary nature to investigate the ability of a human scorer to review wearable EEG and ACM. We first tested the human scorer's ability to independently score the full wearable recordings without access to any machine learning labels. Then, we tested the ability of the scorer to improve the machine learning predictions for the 10% most uncertain areas, when given the full wearable recording and the machine learning predictions for the rest of the recording. We compared the human scorer's results to two automated baselines: one using all the machine learning

[1]The hyperparameter α balances the two factors, and was set to 10, both here and in Chapter 5.

predictions, even the most uncertain ones, and the other interpolating uncertain predictions with surrounding sleep stages. The experimental setup is detailed in Appendix 6.A.

6.2.4 Reliability assessment

The sleep staging accuracy should not be the sole focus for an automated sleep monitoring tool. Indeed, sleep staging serves as a means to an end in a sleep study. The sleep stages are aggregated into sleep parameters, which, along with clinical data, are used to diagnose sleep disorders [146]. We calculated a total of 26 sleep parameters, and compared the sleep parameters predicted on the wearable with the reference parameters of the manually scored PSG, using Pearson's correlation test and a paired t-test.

To further assess the quality of these parameters, we investigated their clinical significance. We computed the Pearson correlation between a few important sleep parameters and the AHI, to verify that the relation between the AHI and automatically estimated sleep stage parameters based on our EEG wearable were consistent with relations found in previous clinical studies on PSG. This experiment was not aimed at proving the diagnostic power of wearable EEG toward sleep apnea diagnosis, as there are more suitable respiration-based devices for this purpose. Instead, it aimed to validate the accuracy of AI-based sleep analysis using wearables to replicate clinical findings from hospital studies. If these parameters are reliable enough, such technology could be employed in home-based sleep studies, for instance, to monitor treatment outcomes long-term. The experimental setup for this experiment is detailed in Appendix 6.A.

6.3 Results

6.3.1 Uncertainty-based training data curation

U-PASS refined the training data and consequently improved the sleep staging performance in two key ways. It first determined the minimal measurements needed based on the amount of uncertainty. As the gyroscope measurements did not contribute to decreasing the data uncertainty, it was removed, and only the cross-head EEG channel and the acceleration data were used. Second, the 1% most ambiguous training samples were removed, improving downstream performance of the trained model. The amount of training data to remove was decided based on the results in Chapter 5. These two steps improved the

sleep staging model with an overall accuracy improvement from 63.9±0.9% to 64.5±0.9% (mean ± standard deviation).

6.3.2 Detecting uncertainty at deployment time

Figure 6.2 shows how uncertainty measures were applied to flag uncertain test data at deployment time, both at the level of whole recordings and at the sample level.

Detecting uncertain recordings

Figure 6.2a demonstrates the recording-wise flagging by plotting the two main factors of the uncertainty measure (6.1) for each recording: the average confidence of the closest training data ($\frac{1}{n}\sum_{i=1}^{n} 1 - c_i$) and the distance to these data ($\frac{1}{n}\sum_{i=1}^{n} d_i$). Abnormal recordings were visually identified as the outliers in the plot (Figure 6.2a), and a decision boundary of 1.216 was applied to the uncertainty metric (6.1) to eliminate these recordings. Alternatively, we could set boundaries based on average confidence (<0.55) and distance (>0.057) to achieve the same result. By rejecting 14% of the data (7 out of 45 recordings), sleep staging accuracy increased from 64.5±0.9% to 68.2±0.9%. The 7 rejected recordings had an average sleep staging accuracy of 43.5±1.8%, confirming that our conservative approach effectively excluded only highly uncertain outliers — a reasonable method given the potential challenges of acquiring new recordings. Similar to the trade-off discussed in Section 5.3.5, the amount of data to reject depends on the trade-off between the cost of a bad recording (and the resulting bad-quality sleep parameters) and the cost of rejecting the recording. These challenges are further discussed in Section 6.4.

This two-factor uncertainty measure not only flagged problematic recordings, but also gave reasons for the uncertainty. The first factor, the distance to the closest training data, identified so-called 'out-of-distribution' data, recordings that deviate significantly from the training data. Such 'anomalous' signals can be caused by technical artefact or atypical brain activity due to biological differences or a pathology. Figure 6.2a, recording 1, demonstrates that the recording with the highest distance to the training data is corrupted by large amounts of noise. The second factor, the confidence of the closest training data, identified recordings with an abnormal amount of epochs close to sleep stage decision boundaries. This entails the presence of lots of sleep stage changes or ambiguous brain activity which can not clearly be assigned to a single sleep stage, both of which are telltales of sleep disorders [146, 147]. Figure 6.2a, recording 2, shows an example with an abnormally high proportion of sleep

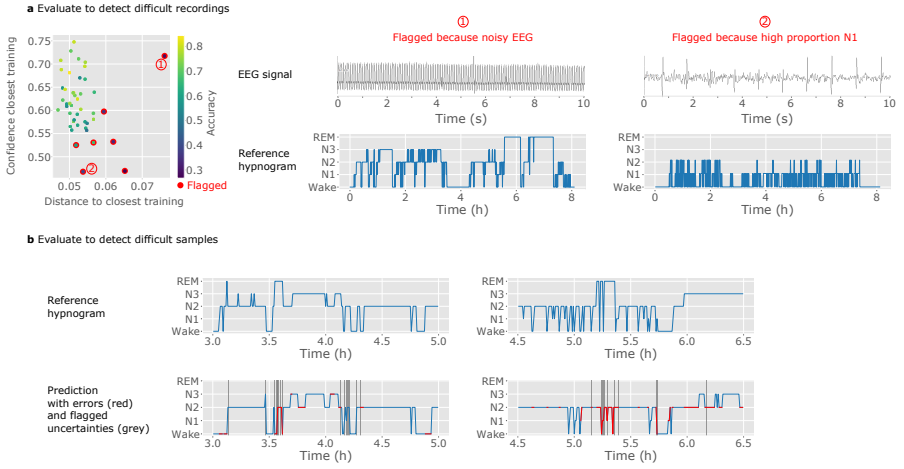


Figure 6.2: U-PASS applied to sleep data captured by a wearable device. (a) In the first evaluation step, recordings with the most prediction uncertainty are rejected as a whole. There can be different reasons for rejecting a recording. The used uncertainty measure takes into account the confidence of the closest training data and the distance from these data. On the left, both factors are visualized for each recording, showing how these influence the accuracy. On the right, two recording snippets and the recordings' hypnograms are shown. The first recording was rejected because of corruption by noise, and the second one was rejected because of the high proportion of sleep stage changes and of the transitory sleep stage N1. (b) In the second evaluation step, uncertainty-based rejection of individual samples can further increase the test accuracy. Samples with uncertainty values higher than the preset threshold were flagged. Two example hypnograms are shown, with the reference sleep stages (top) and the predicted sleep stages (bottom). The samples flagged as uncertain (shown by gray lines) correspond with areas where errors are made (marked in red) and areas with sleep stage changes.

stage changes and sleep stage N1. N1 is a transitory sleep stage, which is notoriously ambiguous and hard to classify [58]. In conclusion, the two factors in this uncertainty metric guide the clinician to decide which action is required. This could be repeating the measurement for that patient, ignoring the recording, replacing the hardware, or asking for a follow-up study at the hospital if the data looks worrisome.

Detecting uncertain samples

After flagging the most uncertain recordings, uncertain segments were flagged in the remaining recordings. With 10% of the data flagged as uncertain, the sleep staging accuracy on the remaining data increased from 68.2±0.9% to 71.2±0.9%. Figure 6.2b shows two segments of data with flagged samples, demonstrating that areas of high uncertainty often corresponded with errors. These errors were primarily attributed to ambiguities arising from sleep stage changes [148] and the sometimes unclear distinction between classes, for example between N2 and REM. The classification of REM-sleep was generally challenging with the cross-head EEG wearable, without EOG or EMG signals available.

6.3.3 Does manual review help for uncertain data?

Figure 6.3 compares our machine learning model to a clinician in five-stage sleep staging on behind-the-ear wearable EEG data. When independently scoring the full wearable recordings, the clinician only achieved 48.7% accuracy, which was much poorer than the SeqSleepNet model's performance of 66.7% accuracy. When we only used the model's 90% most certain sleep stage predictions, it achieved a higher accuracy of 69.5% on that portion of the data. We then tested whether a human scorer could still improve upon the machine learning decisions for the most uncertain areas. Therefore, a human scorer reviewed the incompletely scored wearable data, filling in the blanks by scoring the remaining 10% most uncertain areas of each recording. The total accuracy then became 66.4%, which was still lower than the performance with fully automated scoring. The accuracy achieved by the human scorer was exactly the same as if we would simply interpolate the sleep stage labels using the surrounding sleep stages. We concluded that this human scorer, not trained specifically to interpret wearable behind-the-ear EEG data, could not contribute to achieving more accurate sleep staging with this type of wearable device. Importantly, this was a limited-scale experiment with a single clinician and one wearable device, so caution should be exercised when generalizing these results to all clinicians and all wearable devices. In our experiment, the best sleep staging accuracies were achieved with pure automated scoring, when discarding the most uncertain data based on an uncertainty threshold. However, this would of course result in an incomplete hypnogram, as shown in Figure 6.3.

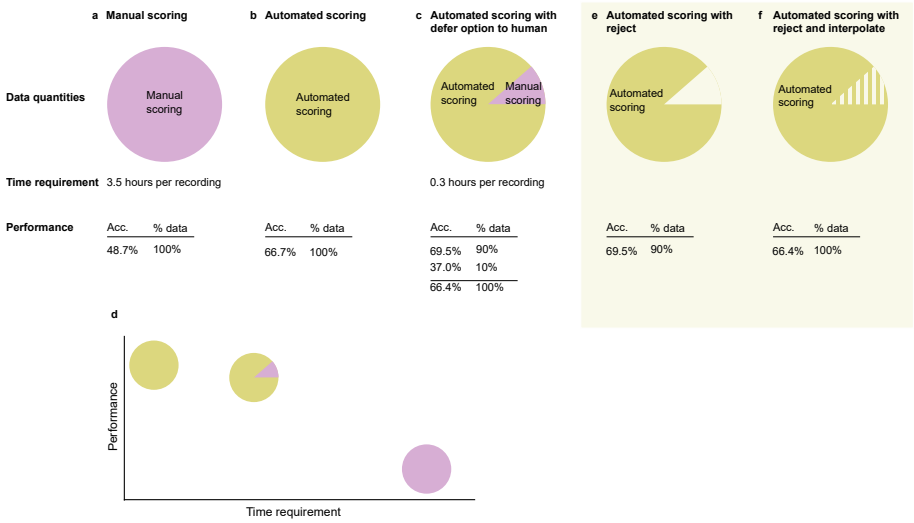


Figure 6.3: Comparison of automated sleep scoring to manual scoring for behind-the-ear wearable EEG recordings. (a) With fully manual scoring, the clinician spent 3.5 hours scoring each recording. (b) With fully automated scoring by our trained sleep model, the performance was higher without requiring any time from a clinician. (c) When only the 10% most uncertain predictions of the model were deferred to and scored by the clinician, the performance was still slightly worse than with fully automated scoring. (d) The performance and time requirement of each scenario is presented visually. (e) When a clinician can not score these uncertain samples better than a machine learning model, what can we do with the uncertain predictions? One option is to simply not predict anything, omitting these most uncertain predictions. When leaving out the 10% most uncertain predictions, the remaining 90% were scored much more accurately. This, however, results in an incomplete hypnogram. (f) Another option is to use simple interpolation of these most uncertain sleep stage labels with the surrounding sleep stages. This approach led to the same accuracy as the review by a clinician, and hence, worse results than taking the original model predictions.

6.3.4 Is automated sleep staging on wearable EEG reliable enough?

As individual sleep parameters are informative for studying long-term sleep evolution and treatment effects, we tested if our machine learning method could reliably predict these parameters based on a one-channel EEG wearable. Based on the predicted sleep stages, we calculated 26 sleep parameters, as shown in Table 6.2. We then computed the correlation of these estimates with the sleep parameters from the manually scored PSG. Uncertain sleep stage predictions were not discarded in this experiment. The reference and prediction were significantly ($p < 0.01$) correlated for 18 out of 26 sleep parameters, demonstrating that trends and inter-patient differences were accurately captured with the wearable and machine learning approach. A paired t-test showed that despite these many significant correlations, the distributions were still significantly different in 17 of the parameters. Hence, although relative differences and trends were well preserved, the absolute values of most of the estimated parameters did not perfectly correspond to the true parameters. In Appendix 6.C, we additionally show the agreement between some of the most important sleep parameters using Bland-Altman plots.

In addition, we also evaluated if the predicted parameters had the same correlations with the AHI as the reference parameters. Figure 6.4 shows the correlation between a few important sleep parameters and the AHI. The automatically predicted and reference sleep parameters are both depicted, with a linear trend line and Pearson's correlation coefficient showing how they correlate with the AHI score. The observed correlations between the sleep parameters and AHI score were largely consistent with previous clinical studies and were preserved by automated estimation (extended results in Table 6.A).

Specifically, three previous studies were found that directly related sleep apnea to sleep architecture parameters [149–152]. The study by Basunia et al. [149] was the only that directly correlated the parameters with the AHI. The study by Redline et al. [152] was conducted with the respiratory disturbance index (RDI), another sleep apnea severity score. The other two studies stratified patients into three groups for a more coarse analysis. Zooming in on a few important parameters, Basunia et al. [149] found that sleep efficiency was negatively correlated with the AHI score, as confirmed by our results (both in the predicted and reference sleep parameters). In the relative times spent in each sleep stage, they found significant correlations for every sleep stage. For N1, we found the same positive correlation with AHI, based on both the predicted and reference sleep stages. For N3, surprisingly, the reference sleep parameters did not find a significant negative correlation contrary to all three previous studies [149–151], but our predicted sleep parameter did. For REM, the

reference and predicted parameters were not significantly negatively correlated with the AHI, unlike in [149]. For other sleep parameters, we did not find conclusive results in the literature, but our predicted and reference parameters agreed on most correlations, like the significant positive correlation of AHI with the amount of awakenings, sleep stage transitions and amount of wake after sleep onset (WASO).

One noteworthy difference between the predicted and reference parameters was the sleep onset as the time to the first N2 segment. The N2 sleep onset was significantly negatively correlated with the AHI score based on the reference sleep stages, but not based on the predicted sleep stages. The results from Basunia et al. [149] agreed with the predicted sleep stages. The fact that the predictions and reference did not entirely agree for this measure was not completely unexpected, given the fact that the sleep onset is hard to reliably estimate, as it is often characterized by multiple transitions between Wake, N1 and N2. Another notable difference between the reference and predicted parameters was the amount of transitions to N3, negatively correlating with the AHI score based on the predicted sleep stages, but not based on the reference sleep stages.

Apart from a few exceptions, we conclude that the machine learning predictions based on wearable EEG resulted in reliable estimates of individual sleep parameters, and that the estimated parameters had the same correlations with the AHI as the the parameters based on the clinician's evaluation of the PSG.

6.4 Discussion

To our knowledge, this is the first time that uncertainty was used to improve the automated analysis of wearable physiological signals. With our novel framework U-PASS, we showed that machine learning can reliably perform sleep staging based on wearable behind-the-ear EEG measurements from elderly patients. We compared the ability of our state-of-the-art machine learning model to a clinician in reading these non-standard physiological data, demonstrating that our machine learning model outperformed the human EEG reader who was not trained to interpret such wearable EEG. The sleep parameters derived from algorithmically determined sleep stages agreed with the reference values derived from the expert-annotated PSG as indicated by Pearson's correlation test. This means that inter-patient differences and trends were well preserved with the wearable AI approach, although the t-test found significant differences in the absolute values of the parameters. We further investigated the potential clinical

Table 6.2: A total of 26 sleep parameters are calculated based on the hypnograms obtained with our deep neural network and the reference hypnograms, as annotated by a clinician based on PSG. Pearson's correlation coefficient (ρ) is calculated between the sleep parameters calculated with both approaches, and the t-test for related samples is performed to identify significant differences between both distributions.

	Predicted[a]	Reference[a]	Pearson's ρ[b]		t-test[c]
			ρ	p-value	p-value
TST (hr)	5.09 ± 1.14	5.50 ± 1.49	**0.91**	0.00	**0.00**
Sleep efficiency (%)	0.65 ± 0.09	0.70 ± 0.14	**0.85**	0.00	**0.00**
W (hr)	2.84 ± 1.10	2.44 ± 1.29	**0.86**	0.00	**0.00**
N1 (hr)	0.09 ± 0.10	0.72 ± 0.63	**0.63**	0.00	**0.00**
N2 (hr)	3.74 ± 0.92	2.72 ± 0.90	**0.55**	0.00	**0.00**
N3 (hr)	0.63 ± 0.42	1.13 ± 0.88	**0.52**	0.00	**0.00**
REM (hr)	0.64 ± 0.47	0.93 ± 0.52	**0.58**	0.00	**0.00**
N1 rel (% TST)	0.02 ± 0.02	0.14 ± 0.13	**0.61**	0.00	**0.00**
N2 rel (% TST)	0.74 ± 0.12	0.51 ± 0.14	0.27	0.10	**0.00**
N3 rel (% TST)	0.12 ± 0.07	0.19 ± 0.13	**0.42**	0.01	**0.00**
REM (% TST)	0.12 ± 0.08	0.16 ± 0.07	**0.53**	0.00	**0.00**
Awakenings	33.76 ± 16.85	36.16 ± 22.89	**0.79**	0.00	0.30
Sleep stage transitions	90.38 ±40.29	135.45 ± 69.69	**0.81**	0.00	**0.00**
Transitions to REM	8.41 ± 4.83	7.87 ± 5.23	0.25	0.13	0.59
Transitions to N3	5.03 ± 2.50	8.11 ± 9.09	0.22	0.18	**0.04**
Time between REM (hr)	0.56 ± 0.29	0.61 ± 0.39	0.14	0.43	0.53
Time between N3 (hr)	0.78 ± 0.40	0.71 ± 0.52	0.16	0.40	0.50
Sleep onset N1 (hr)	1.63 ± 0.67	0.68 ± 0.74	0.30	0.07	**0.00**
Sleep onset N2 (hr)	0.42 ± 0.36	0.62 ± 0.74	**0.63**	0.00	**0.05**
WASO (hr)	1.78 ± 0.86	1.54 ± 1.02	**0.77**	0.00	**0.03**
Awakenings (/TST)	0.06 ± 0.03	0.06 ± 0.04	**0.73**	0.00	0.74
Sleep stage transitions (/TST)	0.15 ± 0.06	0.21 ± 0.12	**0.68**	0.00	**0.00**
Transitions to REM (/TST)	0.01 ± 0.01	0.01 ± 0.01	0.09	0.57	0.45
Transitions to N3 (/TST)	0.01 ± 0.00	0.01 ± 0.01	0.12	0.48	0.10
WASO (/TST)	0.37 ± 0.20	0.32 ± 0.27	**0.71**	0.00	0.12
REM latency (hr)	2.03 ± 1.11	2.08 ± 1.37	**0.74**	0.00	0.76

[a] Represented as mean ± standard deviation.
[b] Significant correlations with $p \leq 0.01$ are marked in bold.
[c] Significant differences with $p \leq 0.05$ are marked in bold.
Abbreviations: WASO: Wake after sleep onset, TST: total sleep time (always expressed in hours), hr: hours.

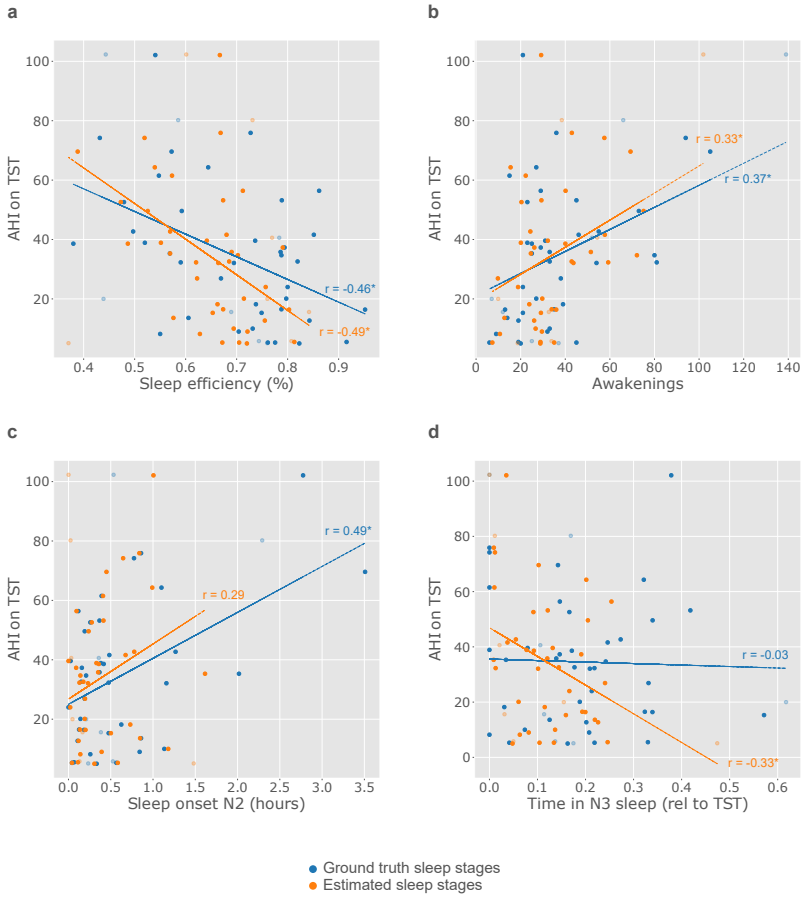


Figure 6.4: Clinical significance of sleep parameters derived from model predictions and PSG reference. The relation between the apnea-hypopnea index on total sleep time (AHI on TST) and key sleep parameters, featuring a linear trend line and Pearson's correlation coefficient (r). Light-colored dots represent recordings discarded due to high uncertainty. The parameters are based on both the reference and predicted sleep stages. Most correlations with AHI are preserved with AI sleep staging, notably for sleep efficiency (a) and number of awakenings (b). Sleep onset (c) and relative time in N3 sleep (d) are two exceptions, with inconsistencies between predicted and reference stages. (Full overview: Table 6.A)

efficacy of the automated pipeline by confirming known correlations of these sleep parameters with a clinical metric, the AHI. We showed that the automated pipeline could indeed reproduce these findings, proving its clinical relevance.

Therefore, relative changes in sleep parameters should be regarded as more reliable indicators than absolute values when wearables and AI are deployed in clinical settings.

U-PASS introduces several improvements to the standard machine learning framework, as extensively discussed in Chapter 5. As in Chapter 5, at training time, we decided which channels to use based on uncertainty and discarded the training data with the most data uncertainty to improve the model. In line with the results from Chapter 5, the optimal performance was achieved when 1% of the training data were discarded. At deployment time, we used the same uncertainty metric as in Chapter 5, a combination of detecting uncertainty related to decision boundaries and out-of-distribution or 'anomalous' data, to flag uncertainties. For the wearable recordings, the higher level of uncertainty and likelihood of unusable data warranted a different approach compared to the one used for PSG in Chapter 5. Instead of just flagging uncertain samples, we now first flagged recordings with overall high uncertainty, followed by more detailed flagging at the sample level. We investigated whether manual review could help to correct these uncertain samples with a clinician-in-the-loop approach, and concluded that this was not true in this particular case. Hence, the active learning and deferral strategies from Chapter 5 were not implemented.

More specifically, our results revealed that a sleep expert was unable to manually correct sleep stages from a wearable device with a behind-the-ear cross-head EEG setup in an elderly population. We have to emphasize that this experiment was limited to a single clinician and one wearable device setup, so we do not have conclusive evidence that this finding can be extrapolated to all sleep experts and all wearables. There is a possibility that sleep scoring on behind-the-ear signals is more difficult to score than from other wearables. However, another study confirmed that machine learning outperforms human scorers on a sleep staging task using around-the-ear EEG, this time in healthy participants [35]. The further complete lack of studies with wearable EEG being scored by clinicians confirms the inherent difficulty. In any case, sleep staging rules are only defined for standard EEG/PSG channels, and so, for any wearable deviating from this standard, the rules would need to be adapted and clinicians would need to be extensively retrained to obtain reliable scoring with minimal interrater variability. We argue that it would be prohibitively costly and time consuming to achieve this for the plethora of wearables that are becoming available.

This is in contrast to sleep staging for PSG recordings, where we (in Chapter 5) and Kang et al. [130] have shown the potential for manually correcting or iteratively improving sleep staging models. Hence, a future research direction for wearables could be to further elevate their signal quality and to rescale and transform the wearable EEG signals to better resemble standard EEG derivations, such that a clinician-in-the-loop approach becomes feasible.

Overall, the major advantage of using our uncertainty-guided framework for remote monitoring applications is the increase in reliability. The accuracy on the test set was improved by an absolute 0.6% through the improvements of the training set. It was then improved by another absolute 3.7% by discarding the seven most problematic recordings of the test set. Using only 90% of the most certain predictions at a segment-level led to another improvement of an absolute 3.0%. Remote monitoring with wearables is prone to noise and variability in data quality, and manual review of data quality in large samples is typically not feasible. Therefore, detecting such uncertainties is crucial to ensure the trustworthiness and usefulness of an automated data processing system. Moreover, automatic sleep staging systems typically perform worse in elderly and diseased populations [41, 153, 154], where sleep disturbances and neurodegenerative disorders result in more fragmented sleep which is difficult to score. We demonstrated that applying uncertainty measures can increase the accuracy in such populations of interest, which is essential to adopt machine learning in clinical practice. In conclusion, frameworks like U-PASS improve the reliability of automated machine learning systems, which is crucial for the success of long-term at-home wearable sleep monitoring systems.

A second advantage of our method, as previously noted in Chapter 5, is its versatility and the fact that it is completely model-agnostic. Since the U-PASS pipeline is compatible with any model that returns output probabilities, we envision its application in various remote monitoring fields, such as seizure detection or cardiac monitoring. The used metrics can also be changed to any other post-hoc uncertainty metric. We decided to use a distance-weighted confidence of neighboring training samples as the uncertainty metric (see Section 6.2.2), but this could be swapped for the Shannon entropy or other uncertainty estimation methods such as Monte Carlo dropout [155]. Depending on the application and needs, the framework could be adapted in multiple ways. Either the recording-wise uncertainty estimation or the sample-wise uncertainty estimation could be excluded, only performing uncertainty estimation on one of the two levels. In some applications, five-class sleep staging may not be needed. For example, if screening for insomnia, the primary parameters of interest could be the total sleep time and time spent awake, requiring only a two-class classification between sleep and wake.

A third benefit of our method is the fact that it can easily be integrated in a clinical workflow with minimal adjustments and requirements. The training requirements of U-PASS do not impact clinical practice, since model training can be done in advance, outside of clinical routine. This ensures minimal disruption to clinical workflows. At test time, U-PASS only requires about 200 MB of working memory to compute the uncertainty measures of a test recording (in case of a training dataset of 50 recordings, and scaling linearly

with the amount of training data). Based on these minimal requirements and the complete model-agnostic nature of our method, U-PASS can feasibly be integrated into any AI-based automated sleep scoring system in daily clinical practice.

A fourth benefit of using our uncertainty-guided framework is that the identified uncertainty in itself can also provide useful information for the diagnosis of sleep disorders. We used an uncertainty metric composed of two factors, with a distance factor detecting out-of-distribution data and a confidence factor detecting closeness to decision boundaries. These two factors give a reason for uncertainty, which can be taken into account to diagnose sleep disorders. For example, we showed that fragmented sleep, typically associated with OSA, leads to more uncertainty as more segments lie on boundaries between different sleep stages (see hypnogram 2 in Figure 6.2a, and Figure 6.A). A low confidence factor in this uncertainty metric could be a relevant marker when screening for OSA. On the other hand, the distance factor will point to atypical, out-of-distribution data. This could be useful for detecting abnormal EEG activity, which can be present in epilepsy or dementia.

We showed that we could identify uncertain predictions such that the remaining ones were more accurate. However, this does not answer the question about what we should do with the identified uncertainties, especially if a clinician may not be able to correct them. As already discussed, the ideal solution we envisage would be to transform the wearable EEG signals to closely resemble standard EEG signals, enabling clinicians to review the uncertain predictions. However, as this solution is not yet available to our best knowledge, uncertainty estimation methods like U-PASS are still, if not even more, invaluable for wearable sleep monitoring. We showed how U-PASS was able to identify bad-quality recordings with overall high uncertainty. Since wearables have the advantage of being cheap recording methods, we would recommend to identify and discard the problematic recordings and repeat these. Consequently, we showed how U-PASS was also able to flag uncertain regions in the hypnogram. Discarding these uncertain predictions led to more accurate sleep staging, but an incomplete hypnogram. If the goal is to estimate sleep parameters, these uncertain predictions should not simply be discarded. As sleep parameters aggregate information from a whole night, small variations in sleep scoring do not significantly change these. Hence, if the uncertainties are limited to small segments and sleep stage transitions, we would recommend to use fully automated sleep staging (the most reliable option as per Figure 6.3) and trust the resulting sleep parameters. On average, these proved to be well aligned with the reference PSG-based sleep parameters (Table 6.2). If large segments of the recording are uncertain, we would again recommend to instead repeat the recording or perform in-hospital PSG to investigate whether the uncertainty

stems from a sleep disorder. If not all sleep parameters are of interest, we should remark that repeating the recording may not be needed, depending on where the uncertainty is concentrated. For example, uncertainty around N2 and N3 sleep is unlikely to affect total sleep time and WASO, and can thus largely be disregarded for these parameters. Therefore, sample-level uncertainty can provide insights into the confidence level of specific sleep parameters. Future research should explore how uncertainty at a sample level can be mapped to confidence estimates for sleep parameters, adding value in the overall diagnostic process.

Two additional practical considerations are worth mentioning when it comes to implementing U-PASS in clinical practice. First, the cost aspect of sleep studies with wearables still needs to be investigated and optimized. Repeating a measurement is the most reliable way to handle a poor-quality recording, but funding schemes may not cover the costs of potential failed sleep studies. One potential strategy to mitigate the likelihood of failed recordings for diagnostic or screening sleep studies using wearables is to provide the patient with the device for an extended period, such as a week, allowing for multiple trials. After each night, uncertainty metrics can serve as direct feedback to the patient, indicating whether the recording was successful or additional recordings are needed. In case of multiple successful recordings, averaging the results of these subsequent measurements may provide a more robust and reliable estimate of sleep, and between-day variability might be informative as well. These cost-related factors should be considered when setting the uncertainty threshold for rejecting a bad-quality recording, balancing the cost of errors against the cost of repeating a study.

A second practical consideration involves effectively communicating uncertainty to clinicians who are not AI experts. To visualize sample-level uncertainty, we recommend displaying the raw signals segmented with the corresponding hypnogram underneath. The hypnogram should be color-coded to indicate the reliability of the predictions: one color for high-quality, reliable predictions and other colors for medium or high uncertainty, similar to the grey areas in Figure 6.2b.

A limitation of the U-PASS pipeline is the need for determining the uncertainty thresholds for marking uncertain regions. These thresholds are subject to adjustment based on specific application demands and precision requirements, as well as the need for scoring completeness. When marking complete recordings as uncertain, the uncertainty threshold can be set lower if it is cheap to replace bad quality recordings by new ones, or when the goal is to detect anomalous recordings with high sensitivity for anomalous brain activity screening. When marking individual samples as uncertain, the threshold depends on the utilization of the resulting hypnogram, and on which sleep parameters need be estimated.

The thresholds can be advantageous in providing the flexibility to deploy this framework in different applications, but domain expertise is needed to choose the thresholds appropriately for every setting.

6.5 Conclusion

In this chapter, we have validated U-PASS, the uncertainty-guided machine learning pipeline for automated sleep staging developed in Chapter 5, for remote monitoring applications with wearable EEG measuring devices. We have shown that improvements can be achieved through both uncertainty-informed refinement of the training data and flagging of uncertain data during deployment.

Our results have implications for the future of AI in sleep monitoring. The vast amount of data that wearable technologies can gather are quite certainly going to surpass the capacities of human review. In this chapter, we showed another problem, which is that sleep experts may not always be able to interpret data from wearables without being specifically trained for this. Therefore, we propose combining AI and wearables to conduct large-scale screenings of populations or for long-term sleep monitoring, particularly to monitor treatment effects. These operations are too costly and time-consuming to be achievable without the help of wearables and AI. We have shown that AI could enable such large-scale operations, by providing fine-grained labeling of the wearable data, and aggregating this information into sleep parameters. These parameters can then be presented in sleep reports for interpretation by human sleep experts. The estimated parameters should be considered approximations of those observed in PSG and must be interpreted with caution. Nevertheless, relative differences between patients and populations are expected to be accurately reflected. Uncertainty estimation plays a crucial role in such an AI-based sleep monitoring approach, as it allows to identify difficult or bad data and avoids potentially wrong decisions and diagnoses. We conclude that an uncertainty-guided framework is essential to guarantee the reliability of AI-based analysis for wearable data in home-based sleep monitoring.

Appendices

6.A Experimental setup

Uncertainty at training and test time

All experiments were performed on the wearable data. The sleep staging model was trained on the full training dataset for 10 training epochs. Early stopping was applied using a validation dataset consisting of 1 to 2 recordings, by retaining the model with the best performance on the validation set. This experiment was repeated 23 times, with different recordings in the validation set for every repeat. Accuracies were computed on the separate test set consisting of 45 recordings. The reported test set accuracies were computed as means ± standard deviations over these 23 repeats.

Manual review

We used ten randomly selected test recordings for the manual review experiment. None of these were marked as difficult to score based on the recording-wise uncertainty detection step. The clinician reported that they were unable to interpret two of these recordings because of too high noise levels. As we did not get manual annotations for these recordings, they were left out of the subsequent analysis. For the remaining eight recordings, we compared one model (randomly selected from the 23 trained models) to the scorings of the clinician.

Estimating sleep parameters

The predicted sleep parameters were calculated for each subject, based on the predicted sleep stage labels. As we trained 23 models, the final sleep parameter for each subject was calculated as the mean over the 23 predicted sleep parameters. The means and standard deviations in Table 6.2 were calculated over all subjects, except the seven recordings that were flagged for high levels of uncertainty. Pearson's correlation test was performed with the subjects as samples. The correlation coefficients in Table 6.A (in Appendix 6.D) were also calculated between the averaged sleep parameters per subject. The same procedure for estimating the sleep parameters was used in the Bland-Altman plots in Figure 6.B (Appendix 6.C) and for the scatter plots in Figure 6.4.

6.B Sources of uncertainty

The uncertainty in sleep staging emerges from two main sources: biological factors such as disease and age, and measurement technique-related factors like segmentation and interference [57]. We can visualize some of the uncertainty-causing factors in the training data. Figure 6.A shows two clinical parameters, the AHI on the total sleep time, and the sleep efficiency (total sleep time over total time in bed) for all the patients in the training set. The AHI is a measure for the sleep apnea severity. Each patient is one dot, and the color of each dot indicates the average confidence over all the samples of that patient's recording. The confidence is the predicted probability for the correct class. It is inversely correlated with both sleep efficiency and the AHI, showing that these clinical factors impact the uncertainty of the model.

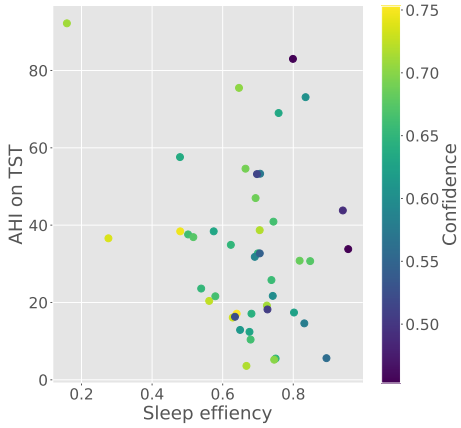


Figure 6.A: Clinical parameters lie at the source of data uncertainty in the wearable training dataset. Mean confidence for every recording in the training set. The confidence is strongly correlated with sleep efficiency (Pearson's $\rho = -0.52, p < 0.05$). It has a weaker, non-significantly relation to the AHI (Pearson's $\rho = -0.12, p = 0.45$).

6.C Bland-Altman plots

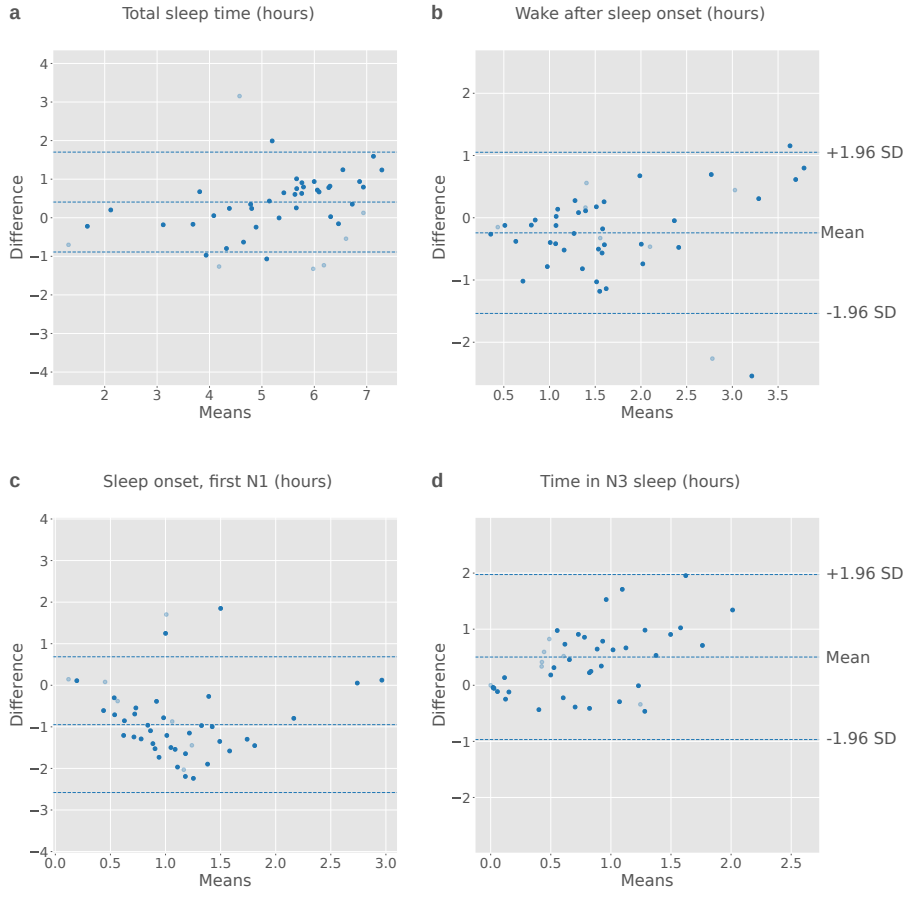


Figure 6.B: Bland-Altman plots for a number of important sleep parameters, where each sample is one patient. Every sample shows an average over the 23 repeats. Each plot shows the agreement between the parameter calculated based on the PSG reference and model predictions. The difference between both is shown on the y-axis, and the mean of the two measures is plotted on the x-axis. The light-colored dots represent recordings that were discarded because of high levels of uncertainty. These Bland-Altman plots show an overall good agreement between the predicted and reference sleep parameters.

6.D Sleep parameters' correlations with AHI

Table 6.A demonstrates the clinical significance of the sleep parameters. It shows the Pearson's correlations between the 26 sleep parameters from Table 6.2

and the AHI score, and compares these results to findings from prior studies.

Table 6.A: A total of 26 sleep parameters calculated based on the hypnograms obtained with our deep neural network and the reference hypnograms, as annotated by a clinician based on PSG. Pearson's correlation coefficient (ρ) is calculated between the sleep parameters and the AHI on total sleep time (TST).

	Pearson's ρ ($\mathbf{p \leq 0.05}$)				Clinical papers	
	Predicted		Reference		Sign.	Non-sign.
	ρ	p-value	ρ	p-value		
TST (hr)	-0.10	0.55	-0.16	0.33	[149, 150]: -	[151]
Sleep efficiency (%)	**-0.49**	0.00	**-0.46**	0.00	[149, 152]: -	[150, 151]
W (hr)	**0.48**	0.00	**0.51**	0.00		
N1 (hr)	**0.40**	0.01	**0.44**	0.01		
N2 (hr)	0.05	0.75	**-0.32**	0.05		
N3 (hr)	**-0.34**	0.04	-0.14	0.41		
REM (hr)	-0.12	0.46	-0.21	0.20	[150]: -	
N1 rel (% TST)	**0.49**	0.00	**0.44**	0.01	[149, 151, 152]: +	[150]
N2 rel (% TST)	0.20	0.22	-0.27	0.10	[150, 152]: +, [149]: -	[151]
N3 rel (% TST)	**-0.33**	0.04	-0.03	0.85	[149–152]: -	
REM (% TST)	-0.11	0.52	-0.18	0.27	[149, 152]: -	[151]
Awakenings	**0.33**	0.04	**0.37**	0.02		
Sleep stage transitions	**0.32**	0.05	**0.37**	0.02		
Transitions to REM	0.00	1.00	0.01	0.93		
Transitions to N3	**-0.32**	0.05	-0.03	0.88		
Time between REM (hr)	0.12	0.51	**0.36**	0.04		
Time between N3 (hr)	-0.14	0.46	-0.07	0.72		
Sleep onset N1 (hr)	0.26	0.12	0.31	0.06		[149]
Sleep onset N2 (hr)	0.29	0.08	**0.49**	0.00		[149]
WASO (hr)	**0.41**	0.01	**0.35**	0.03		
Awakenings (/TST)	**0.42**	0.01	**0.43**	0.01	[149, 150, 152] *	
Sleep stage transitions (/TST)	**0.44**	0.01	**0.44**	0.01		
Transitions to REM (/TST)	0.04	0.80	0.06	0.70		
Transitions to N3 (/TST)	-0.32	0.05	0.05	0.76		
WASO (/TST)	**0.43**	0.01	**0.34**	0.03		
REM latency (hr)	0.06	0.74	0.18	0.30	[150]: +	

Significant correlations ($p \leq 0.05$) are marked in bold.
Studies that found a significant correlation are reported in the 'sign.' column, with a '+' or '-' indicating the type of correlation, and papers that found the correlation was non-significant correlation are reported in the 'non-sign.' column.
References: [149–152]
Abbreviations: WASO: Wake after sleep onset, hr: hours, +: positive correlation, -: negative correlation, *: positive correlation with the arousal index, which is strongly related to but not equal to the amount of wakenings over TST.

Part IV

Validation in clinical populations

7 | Automated sleep staging in clinical practice

This chapter is based on segments from

- L. Micca, **E. R. M. Heremans**, D. Testelmans, P. Borzée, W. Vandenberghe, M. De Vos, and M. Gilat, "AI-driven automated sleep staging in Parkinson's disease", *27th Congress of the European Sleep Research Society*, Seville, Spain, 2024.
- R. Proost, **E. R. M. Heremans**, L. Lagae, W. Van Paesschen, M. De Vos, and K. Jansen, "Automated sleep staging on reduced channels in children with epilepsy," *Frontiers in Neurology*, vol. 15, pp. 1390465, 2024.
- J. Macea Ortiz, **E. R.M. Heremans**, R. Proost, M. De Vos, and W. Van Paesschen, "Automated sleep staging using deep learning on wearable EEG and accelerometry in people living with epilepsy," Submitted for review, 2024.
- L. Van den Bulcke, H. Davidoff, **E. R. M. Heremans**, Y. Potts, K. Vansteelandt, M. De Vos, D. Christiaens, L. Emsell, L. H. Jacobson, D. Hoyer, B. Buyse, M. Vandenbulcke, D. Testelmans, and M. Van Den Bossche, "Acoustic stimulation to improve slow-wave sleep in Alzheimer's disease: a multiple night at-home intervention," *The American Journal of Geriatric Psychiatry*, 2024.
- L. Van den Bulcke, A. M. Peeters, **E. R. M. Heremans**, H. Davidoff, P. Borzée, M. De Vos, L. Emsell, J. Van den Stock, M. De Roo, J. Tournoy, B. Buyse, M. Vandenbulcke, C. Van Audenhove, D. Testelmans, and M. Van Den Bossche, "Acoustic stimulation as a promising technique to enhance slow-wave sleep in Alzheimer's disease: results of a pilot study," *Journal of Clinical Sleep Medicine*, vol. 19, no. 12, pp. 2107–2112, 2022.

ABSTRACT | To demonstrate the clinical utility of automated sleep analysis, it is essential to validate its effectiveness in relevant populations. Using AI tools to track sleep can significantly enhance our understanding of its relationship with various conditions. This approach may help identify potential sleep biomarkers and therapeutic targets for different disorders. With the support of clinical collaborators from the Group of Biomedical Sciences at KU Leuven, this chapter explores how the automated sleep staging tools developed in Chapter 3 and Chapter 4 can help to quantify sleep and correlate its characteristics with other clinical parameters. It highlights the potential of automated sleep staging for individuals with Parkinson's disease, Alzheimer's disease, and pediatric and adult epilepsy patients.

7.1 Introduction

Quantifying sleep with AI can significantly help in understanding the role it plays in various conditions. Automated sleep analysis has the potential to identify subtle changes in sleep architecture that might be indicative of underlying health issues. This chapter focuses on validating the transfer learning tools discussed and developed in Chapter 3-4 for characterizing sleep in clinical populations. Through these evaluations in different datasets, the role of automated sleep analysis is explored in different clinical populations and applications.

To validate the developed sleep staging methods in clinical settings, this chapter introduces new datasets that are not utilized elsewhere in this thesis and are therefore not described in Chapter 2. These datasets, detailed in Table 7.1, include individuals with Parkinson's disease (PD), AD, and epilepsy. Each dataset has unique characteristics. In the PD study, PSG data was fully labeled with manual sleep staging. The pediatric epilepsy dataset also consisted of manually labeled PSG data and was divided into three clinical subgroups. For the adult epilepsy study, there was a limited set of manually labeled simultaneous PSG and wearable device measurements, along with a large set of unlabeled wearable measurements. Finally, the AD DREEM2 dataset, different from the SDAD dataset from Chapter 2, was collected with unlabeled wearable device measurements. Due to the distinct characteristics of these four datasets, each required a tailored approach to sleep staging and result analysis.

Transfer learning methods are primarily aimed at boosting sleep staging accuracies, with different approaches applying to different scenarios. For labeled PSG datasets, fine-tuning a neural network pre-trained on another PSG dataset is generally effective. If the PSG dataset is sufficiently large, supervised training from scratch without any transfer learning is a potential approach as well. For wearable datasets, if labeled training data are available from simultaneous PSG recordings, the feature matching approach is preferable to fine-tuning, as it leverages information from paired PSG and wearable measurements. When dealing with large amounts of unlabeled data, semi-supervised or unsupervised domain adaptation approaches should be used, depending on the availability of labeled data. Table 7.2 shows these different methods and scenarios in which they should be used.

Next to the different approaches to sleep staging, the unique characteristics of the different datasets also allowed for other types of analysis. Accordingly, we also analyzed a few interesting between-group differences, correlations, and relationships in the sleep and patient characteristics and sleep staging performances.

Table 7.1: Summary of the datasets used for clinical validation in this chapter.

Dataset	Nb. recordings (patients)	Subject characteristics	PSG	Wearable	Collaboration
Parkinson's disease	46 40	PD Healthy older adults	In-laboratory	/	Micca et al
Pediatric epilepsy	47 74 55	Healthy, age 4-18 WCE, age 4-18 DRE, age 4-18	In-laboratory	/	Proost et al
Adult epilepsy	20 203 (50)	Adult epilepsy patients	In-laboratory /	Behind-the-ear Behind-the-ear	Macea Ortiz et al
Alzheimer's disease*	217 (15)	AD, mean age 76	/	Headband	Van den Bulcke et al

* Not the SDAD dataset.

Table 7.2: Transfer learning approaches and the ideal scenarios in which to use them.

Transfer learning approach	Labels needed	PSG	Wearable
Fine-tuning	✓	✓	
Feature matching	✓	✓	✓
Semi-supervised domain adaptation	Partly	Partly	✓
Unsupervised domain adaptation			✓

7.2 Parkinson's disease

People with PD present sleep alterations that can increase uncertainty during manual sleep staging on PSG [57, 156]. Moreover, manual sleep staging is a time-consuming task with an interrater reliability that is only moderate in PD [157, 158]. As the efficacy of automated sleep staging in PD is unknown, we evaluated SeqSleepNet's performance in this context to potentially reduce the time required for sleep staging.

7.2.1 Dataset

Forty-six people with PD and 40 healthy older adults (cognitively intact elderly or CIE) underwent overnight PSG at the University Hospitals Leuven (Belgium). The PSG was scored offline in 30-second epochs and verified by a certified sleep technologist.

7.2.2 Methods

Sleep staging The raw PSG data was automatically labeled using SeqSleepNet, pre-trained on the MASS dataset (see Section 2.1.1), using five PSG channels: C4-A1, F4-A1, O2-A1, EOG, and chin EMG. As the PD dataset consists of labeled PSG, **fine-tuning** was used to adapt the algorithm to the PD dataset (see Table 7.2). The fine-tuned version was evaluated with 10-fold cross-validation.

Analysis Spearman correlations were explored, comparing accuracy of both algorithm versions to sleep-related potentially confounding factors (i.e., arousal index and periodic-limbs movements during REM sleep).

7.2.3 Results

Sleep staging The pre-trained algorithm performed with 79% accuracy for the CIE ($\varkappa$=0.71) and 75% for PD ($\varkappa$=0.65). Fine-tuning of the algorithm led to an increase in sleep-staging performance (86% and $\varkappa$=0.80 for CIE and 82% and $\varkappa$=0.73 for PD). The fine-tuned algorithm exhibited strong sensitivity above 80% for Stages Wake, N2 and REM in both populations (PD-Wake=93%, PD-N2=85%, PD-REM=84%; CIE-Wake=94%, CIE-N2=87%, CIE-REM=93%). However, sensitivity for Stage N3 was lower in PD (73%) than in CIE (80%), and Stage N1 demonstrated poor sensitivity in both populations, with values of 31% for PD and 35% for CIE.

Analysis No correlations survived Bonferroni correction.

7.3 Epilepsy in pediatric population

The bidirectional relationship between sleep and epilepsy has been a longstanding area of interest, and a growing body of research on this topic focuses on the pediatric population [159–161]. To study the complex interactions between sleep and epilepsy-related factors on a large scale, we need to investigate the efficacy of automated sleep staging tools.

7.3.1 Dataset

In-hospital overnight video-EEG, along with EOG and chin EMG, was recorded in 176 children between 4 and 18 years of age: 47 children without epilepsy, 74 with WCE, and 55 with DRE. Manual sleep staging was performed on all these recordings by a certified sleep technologist.

7.3.2 Methods

Sleep staging SeqSleepNet was used to perform automated sleep staging using three channels: C4-A1, EOG, and chin EMG. It was first pre-trained on the

MASS dataset. Then, in line with the reasoning outlined in Table 7.2, it was **fine-tuned** on the pediatric population, using 15-fold cross-validation.

Analysis First, between-group differences in sleep staging performance were analysed with univariate analyses conducted using the Mann–Whitney U test. Second, between-group differences in the amount of REM-sleep were analysed, since prior literature has found decreased REM sleep in children with DRE [159].

7.3.3 Results

Sleep staging Table 7.3 shows the five-class sleep staging accuracy of the automatic sleep staging algorithm. Fine-tuning to the pediatric population increases the accuracies to 84.7% for the children without epilepsy, 83.5% for those with WCE, and 80.8% for those with DRE. Overall performance per sleep stage was assessed, with an F1-score of 0.91 for Wake, 0.50 for N1, 0.83 for N2, 0.84 for N3, and 0.86 for REM.

Table 7.3: Sleep staging performance of SeqSleepNet on pediatric epilepsy population.

	Pre-trained		Fine-tuned	
	Accuracy	κ	Accuracy	κ
No epilepsy	0.781 ± 0.013	0.697 ± 0.018	0.847 ± 0.012	0.789 ± 0.017
Well controlled epilepsy	0.749 ± 0.015	0.653 ±0.020	0.835 ± 0.010	0.769 ± 0.014
Drug-resistant epilepsy	0.638 ± 0.027	0.531 ± 0.030	0.808 ± 0.014	0.726 ± 0.019

Analysis The reasons for the lower sleep staging performance in the DRE group were investigated. First, there was a trend towards more N1 in DRE patients, although this difference was non-significant. Second, in the DRE group, there were significantly more patients with intellectual disability and multifocal and/or abundant interictal epileptiform discharges (IEDs). All three of these conditions were related with lower κ-scores in the complete patient cohort: the Mann–Whitney U test showed significant differences with $p < 0.001$ for intellectual disability, $p = 0.013$ for multifocal IEDs, and $p < 0.001$ for abundant IEDs.

In children with DRE, the amount of REM sleep time was significantly lower than in children with WCE and without epilepsy. This finding was reliably reproduced with automated scoring, as seen in Figure 7.1.

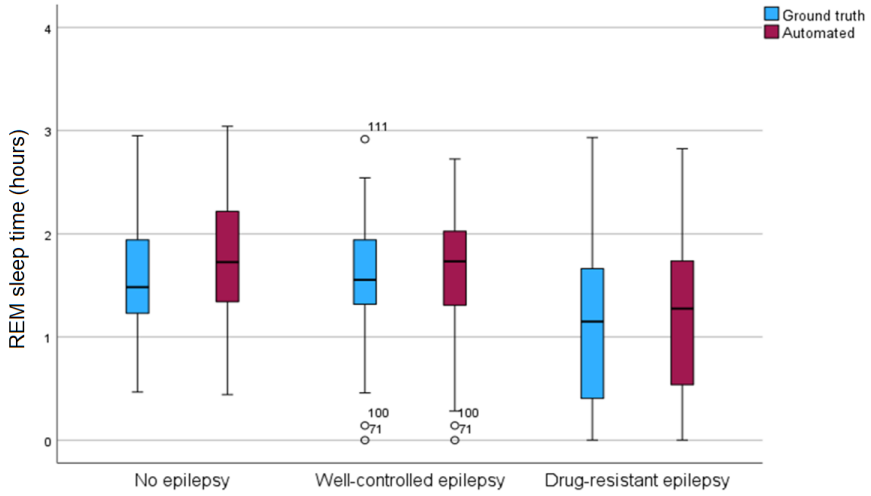


Figure 7.1: Box-plot comparing REM sleep time according to ground truth scoring and scoring by the fine-tuned algorithm.

7.4 Epilepsy in adult population

As discussed in Section 7.3, automated sleep staging tools can facilitate studying the complex interactions between sleep and epilepsy. However, achieving this on a large scale necessitates the use of wearable devices. In this section, we focus on sleep staging in an adult epilepsy population, making use of wearable device recordings.

7.4.1 Dataset

A total of 50 patients of the SeizeIT2 study [31] were included in this sleep assessment study. These participants were each monitored during at least three nights using simultaneous recordings of 25-channel EEG and SensorDot measurements. A certified clinician performed sleep staging for twenty of the recordings, using the 25-channel EEG. The final dataset consisted of 223 recordings in total, with 3-5 recordings per patient (4.46 on average).

7.4.2 Methods

Sleep staging SeqSleepNet performed automated sleep staging on the 25-channel EEG, using the three EEG channels C4-P9, F4-P9, and O2-P9, and on the SensorDot data, using the cross-head EEG and the ACM channel. The model was pre-trained on sleep records from the CIE individuals of the SDAD dataset with simultaneous EEG and SensorDot recordings (see Section 2.1.2). Then, we analyzed the potential to further improve the model by adapting it to recordings of individual patients of the epilepsy dataset. The limited labeled data (a maximum of one recording per patient) was needed to test the performance of the sleep staging model, and the further absence of labeled data warranted an **unsupervised domain adaptation** approach (see Table 7.2). Two personalization approaches were tested: recording-wise adaptation, as described in Chapter 4, and patient-wise adaptation, in which all recordings from an individual patient were used. The pre-trained and personalized models were evaluated using the 20 labeled recordings.

Analysis This dataset is characterized by a large number of unlabeled recordings, which presents a challenge when trying to objectively assess the sleep staging performance. Hence, instead of quantifying the performance, we quantify the agreement between sleep parameters estimated from the standard EEG and from the wearable device recordings. Sleep is characterized using 28 sleep parameters estimated from the sleep stages for all 223 recordings. The concordance between the two sets of parameters is measured through their Spearman correlation coefficient.

7.4.3 Results

Sleep staging Table 7.4 shows the sleep staging results, evaluated on the 20 labeled recordings. The recording-wise adaptation performs best for the recordings from the wearable, whereas the pre-trained network performs best for the standard EEG recordings. Strikingly, using multiple recordings per patient to perform personalization to that patient does not improve the results.

Analysis The sleep parameters computed from the sleep stages are shown in Table 7.5 for all 223 recordings. The Spearman correlation test finds significant correlations between AI-derived sleep parameters based on standard EEG and based on the wearable device in all but three parameters.

Table 7.4: Sleep staging performance of SeqSleepNet on adult epilepsy population.

	Accuracy	κ
Standard EEG		
Pre-trained	**0.728 ± 0.017**	**0.592 ± 0.024**
Recording-wise adaptation	0.717 ± 0.018	0.581 ± 0.025
Subject-wise adaptation	0.717 ± 0.018	0.580 ± 0.025
Wearable		
Pre-trained	0.613 ± 0.027	0.433 ± 0.038
Recording-wise adaptation	**0.631 ± 0.025**	**0.455 ± 0.036**
Subject-wise adaptation	0.614 ± 0.029	0.441 ± 0.037

Results are reported as mean ± standard error.
The best result per recording device is marked in bold.

7.5 Alzheimer's disease

The previous sections explore the potential for sleep staging to characterize and study sleep in relevant clinical populations. This last section goes beyond characterization, and investigates sleep staging for therapeutic interventions.

A reduction in slow wave activity (SWA) is a key feature of AD and has been linked to cognitive decline and to the pathophysiology of AD [162,163]. Acoustic stimulation of slow oscillations during N3 sleep has been shown to enhance slow wave activity (SWA) and memory consolidation in one-night experiments on healthy participants [164–166]. Therefore, we aimed to investigate the impact of auditory closed-loop stimulation during N3 sleep on SWA in patients with AD over several nights in their home environment. In order to objectively assess the benefits on SWA during N3 sleep, reliable detection of N3 sleep is essential. This is where automated sleep staging methods come into play.

7.5.1 Dataset

We present an at-home trial in 15 older patients with AD (mean age: 76.27 (SD=6.06), mean MOCA-score: 16.07 (SD=6.94), 6 women). We assess the effect of stimulation on SWA and SWS in our two-week mobile closed-loop auditory stimulation intervention using the commercially available wearable headband device (DREEM2, Rythm SAS, Paris). Patients wore the device for an initial two nights without stimulation to establish a baseline, followed by 14

Table 7.5: Sleep parameters estimated from the wearable and from the 25-channel EEG, and Spearman correlation between both sets of parameters.

	SensorDot	EEG	Spearman ρ	
			ρ	p-value
TST	5.46 ± 1.43	5.86 ± 1.50	0.93	**0.00**
SleepEff	0.80 ± 0.07	0.86 ± 0.08	0.46	**0.00**
#W	1.35 ± 0.56	0.95 ± 0.63	0.58	**0.00**
#N2	3.04 ± 1.08	3.02 ± 1.20	0.82	**0.00**
#N3	0.87 ± 0.47	1.09 ± 0.50	0.69	**0.00**
#R	1.56 ± 0.62	1.72 ± 0.58	0.64	**0.00**
Rel#N2	0.55 ± 0.12	0.50 ± 0.11	0.60	**0.00**
Rel#N3	0.16 ± 0.08	0.19 ± 0.08	0.68	**0.00**
Rel#R	0.29 ± 0.09	0.30 ± 0.07	0.41	**0.00**
Awakenings	24.43 ± 9.18	21.51 ± 10.24	0.46	**0.00**
SlStCh	78.77 ± 29.11	71.86 ± 26.39	0.47	**0.00**
ChToR	16.87 ± 8.60	14.17 ± 7.71	0.46	**0.00**
ChToN3	7.64 ± 4.77	8.68 ± 4.40	0.40	**0.00**
TimeBetweenW	0.24 ± 0.09	0.33 ± 0.20	0.18	**0.01**
TimeBetweenR	0.33 ± 0.18	0.40 ± 0.25	0.23	**0.00**
TimeBetweenN3	0.63 ± 0.48	0.48 ± 0.45	0.40	**0.00**
WBoutDuration	0.06 ± 0.03	0.04 ± 0.03	0.50	**0.00**
RBoutDuration	0.11 ± 0.06	0.16 ± 0.10	0.26	**0.00**
N3BoutDuration	0.15 ± 0.11	0.15 ± 0.09	0.51	**0.00**
SleepOnsetN1	0.11 ± 0.17	1.94 ± 2.02	0.05	0.43
SleepOnsetN2	0.11 ± 0.17	0.11 ± 0.15	0.57	**0.00**
WASO	1.24 ± 0.55	0.62 ± 0.54	0.46	**0.00**
RelAwakenings	0.04 ± 0.01	0.03 ± 0.01	0.19	**0.00**
RelSlStCh	0.12 ± 0.03	0.10 ± 0.03	0.15	0.03
RelChToR	0.03 ± 0.01	0.02 ± 0.01	0.24	**0.00**
RelChToN3	0.01 ± 0.01	0.01 ± 0.01	0.37	**0.00**
RelWakeTime	0.24 ± 0.11	0.11 ± 0.10	0.36	**0.00**
RemLatency	0.98 ± 0.84	0.78 ± 0.69	0.04	0.50

Parameters are reported as mean ± standard deviation.
The Spearman correlation test is performed with the 223 recordings as samples.
Significant correlations with $p \leq 0.01$ are marked in bold.
Abbreviations of the sleep parameters (hypnogram features) are explained in Appendix 8.A.

consecutive nights with stimulation enabled (Figure 7.2). The lack of PSG in this dataset implies there are no ground truth sleep stage labels.



Figure 7.2: Representation of the DREEM2 device and the experimental procedure. A. The DREEM2 device, self-applied by the participants. B. The study protocol: patients first wore the device for two consecutive nights with the feature for stimulation turned off to establish a baseline (OFF). Then, they wore the headband for 14 consecutive 'stimulation nights', in which stimulation (STIM) or no stimulation (SHAM) was delivered when SWS was detected.

The DREEM2 headband records data from five EEG electrodes, an accelerometer, and an oximeter, and uses its own machine learning algorithm to perform automated sleep staging. It performs acoustic stimulation through bone conduction during N3 sleep, each time delivering a train of two consecutive 45 dB stimulations.

7.5.2 Methods

Sleep staging Although DREEM2 has its own sleep staging algorithm, we need to quantify N3 sleep independently in order to objectively assess the benefits of its SWA stimulation. We independently score the raw EEG data using SeqSleepNet (Section 2.2.3). We pre-train the algorithm on the public Dreem - Headband dataset (Section 2.1.1), recorded from healthy adults with an earlier version of the same device. Then, we use **unsupervised adversarial domain adaptation** (Chapter 4) to adapt the algorithm to individual patients, this choice being motivated by the complete lack of labeled data (Table 7.2).

Analysis This dataset lacks ground truth scoring, so we are not able to objectively assess the sleep staging performance. The only way of establishing trust is to assess whether the SeqSleepNet and DREEM2 scoring show similar results, and through visual verification by a human scorer. Hence, in a small pilot study on a subset of seven participants, we establish whether sleep parameters from DREEM2 and derived from SeqSleepNet scoring show similar results. The SeqSleepNet scoring is verified and corrected by three human annotators to achieve a human-machine 'consensus' scoring.

Table 7.6: Sleep macrostructure for nights 2 and 16. The results as shown represent respectively those by the embedded automatic algorithm of the DREEM2 headband and those of the scoring of SeqSleepNet with expert verification.

	DREEM2 scoring			**SeqSleepNet scoring**		
	Night 2	Night 16	p	Night 2	Night 16	p
Total sleep time	397.43±79.39	421.29±65.23	0.394	468.18±82.73	473.91±70.39	0.774
Sleep onset latency (min)	41.07±40.52	30.64±38.47	0.236	12.32±4.93	9.30±3.71	0.093
Wake after sleep onset (min)	116.79±45.19	88.86±73.17	0.237	77.68±43.43	60.34±28.36	0.275
Stage N3 (min)	35.07±38.67	68.64±46.64	0.019	40.86±27.28	66.93±15.28	0.053
Stage N3 (%)	8.76±9.98	15.80±9.90	0.054	8.83±4.92	14.35±3.15	0.034

Sleep parameters are reported as mean ± standard deviation.
Two-sided p-values are mentioned of the paired t-tests / of the non-parametric Wilcoxon signed rank test for the DREEM2 results of sleep onset latency and wake after sleep onset.

After trust is established through this first analysis, a second analysis, most important to this study, is to evaluate whether SWA indeed increases as a result of the acoustic stimulation. This is evaluated by combining the DREEM2 and SeqSleepNet scoring.

7.5.3 Results

Analysis In the initial pilot study, we establish that the changes in sleep macrostructure are consistent between manually verified SeqSleepNet scoring and DREEM2 scoring (Table 7.6). Although the absolute values of the parameters differ, the relative changes and p-values are consistent between the two approaches.

When combining the DREEM2 and SeqSleepNet scoring, Figure 7.3 shows that stimulation during SWS significantly enhances SWA. It also increases time in SWS at a group level, when comparing the amount of SWS between night 2 and 16.

7.6 Discussion

This chapter discussed sleep staging results for two PSG datasets and two datasets acquired with different wearable EEG monitoring devices. Tailored transfer learning approaches were applied to both types of datasets depending on the availability of labels: supervised fine-tuning was used to optimize sleep

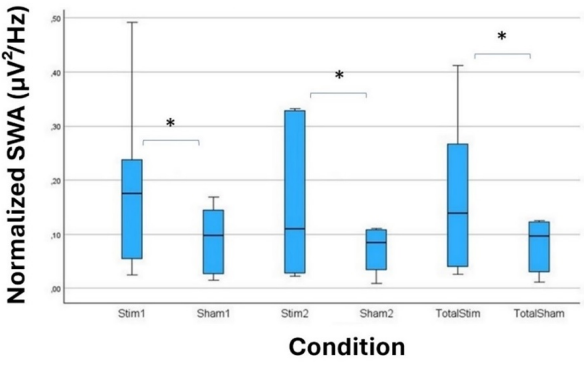


Figure 7.3: Normalized power in the slow wave frequency range of the 4-second window following the first stimulation/sham, the second stimulation/sham, and the total stimulations/shams. We only include data from trains of two stimulations at 45 dB and shams that occurred during epochs identified as SWS by both DREEM2 and the scoring by SeqSleepNet.

staging performance for the labeled PSG datasets, while unsupervised domain adaptation was employed for the unlabeled wearable datasets.

SeqSleepNet demonstrated clinical expert-level performance for sleep staging on PSG in both the PD dataset and the pediatric epilepsy dataset following fine-tuning to the respective populations. It performed particularly well in CIE adults and children without epilepsy and with WCE. Performance in adults with PD and children with DRE was acceptable, although it was significantly lower, indicating that disease is a significant source of uncertainty in sleep staging. In PD patients, the sensitivity for N3 was significantly reduced compared to CIE. For the DRE group, the lower accuracy could be explained by more time spent in N1 and by abundant interictal epileptiform discharges and intellectual disability leading to less recognizable sleep stages. These disease-related challenges are not unique to AI-based sleep staging, as humans also exhibit lower inter-rater agreement in diseased populations [57, 59, 60]. The expert-level accuracies obtained in both the PD and pediatric epilepsy population underscore the potential use of automated sleep staging approaches in daily clinical practice. For instance, the significant decrease in REM sleep time in children with DRE was reliably detected with automated scoring, demonstrating its diagnostic potential.

Wearable devices provide a more practical and scalable solution than PSG for large-scale studies of sleep in relation to different clinical conditions. This chapter introduced two (largely) unlabeled datasets acquired with wearable EEG

devices, using unsupervised adversarial domain adaptation to adapt SeqSleepNet to these datasets. In the adult epilepsy dataset, a few labeled recordings were used to evaluate the performance of the sleep staging network, showing results comparable to those obtained with wearables in other clinical cohorts. Moreover, the majority of sleep parameters derived from wearable and standard EEG data exhibited strong correlations, indicating that trends in sleep statistics were preserved despite variations in sleep staging accuracy. This finding aligns with the conclusions in Chapter 6, highlighting the clinical potential of wearable EEG devices. In the Alzheimer's dataset recorded with the DREEM2 device, the lack of labels or another recording device prevented objective assessment of sleep staging performance. Comparing two automated sleep staging methods (the DREEM2 algorithm and SeqSleepNet), both showed the same relative changes in sleep structure due to the intervention. Through acoustic stimulation during N3 sleep, SWA was enhanced, and participants spent more time in N3 sleep. These findings demonstrate that machine learning tools for sleep analysis are not only useful for diagnostic purposes but also have potential for therapeutic applications through real-time sleep staging and active intervention in the sleep cycle.

The results on the adult epilepsy dataset warrant some additional discussion from a technical perspective. Unsupervised adaptation to individual recordings improved performance for the wearable device but did not improve sleep staging for the standard EEG setup. This could be attributed to the lower baseline performance for the wearable device and the higher levels of noise and variability in the wearable device recordings (see Section 1.2.2), making recording-wise adaptation more beneficial.

In domain adaptation terms, each patient could be viewed as an individual domain with distinct sleep and EEG characteristics. However, in the adult epilepsy dataset, patient-wise domain adaptation did not outperform recording-wise adaptation, suggesting that differences between individual recordings may be greater than differences between patients in this dataset. Future research should explore how many recordings of a single patient are needed before unsupervised patient-wise adaptation becomes truly useful[1]. Further investigation is required to determine whether unsupervised adaptation is best performed at the population level, subject level, or recording level. The optimal approach may vary depending on factors such as inter-individual differences and night-to-night variability within specific populations, and the volume of data that can be collected per subject.

Different problems and datasets require different transfer learning approaches.

[1] K. P. Lorenzen, **E. R. M. Heremans**, M. De Vos, and K. B. Mikkelsen, "Personalization of automatic sleep scoring: How best to adapt models to personal domains in wearable EEG," Accepted in *IEEE Journal of Biomedical and Health Informatics*, 2024.

For PSG datasets, sufficient labeled data are typically available to adopt a fine-tuning approach, transforming a 'base' model into a specialized model tailored to each sleep laboratory and/or specific patient population. For sleep staging using wearable devices, semi-supervised or unsupervised domain adaptation approaches offer flexibility and scalability, allowing to adapt a model with little to no labeled data and to leverage large amounts of unlabeled data. Such approaches can be useful to adapt a wearable-based sleep staging model to new patient populations, subjects, or recordings. However, it must be acknowledged that unsupervised domain adaptation typically achieves only limited accuracy improvements. For scenarios with a large domain mismatch, like the adaptation of a PSG-based sleep staging model to a new wearable device, a supervised learning approach like feature matching is recommended.

Having discussed the main findings of this chapter, let us reflect on whether sleep staging models are ready for use in clinical practice. For the PSG datasets, the expert-level accuracies following the fine-tuning of a sleep staging model show that appropriately trained neural networks achieve adequate performances for clinical use. In the case of wearable devices, this chapter demonstrated that using a model trained on data from a different population acquired with the same wearable, unsupervised learning can improve sleep staging accuracies for the new population. The resulting sleep stages were sufficiently accurate to track sleep parameters in a population and detect changes in sleep over time. This shows that automated sleep staging using wearable devices can already be deployed for long-term monitoring of large populations. Overall, the results on both PSG and wearable devices demonstrate that by employing transfer learning approaches, we can achieve adequate sleep staging accuracy for the clinical application of automated sleep staging models.

Despite the demonstrated adequate performance of sleep staging models in both wearable and PSG datasets, these machine learning tools should not be blindly trusted. Average accuracy measures do not ensure consistent performance across all individual recordings. Therefore, methods like U-PASS (Chapter 5-6) which identify uncertainty and integrate clinician feedback into the machine learning pipeline, remain crucial for the successful implementation of these methods in clinical practice.

7.7 Conclusion

With various applications on different patient populations, this chapter demonstrates the validity of automated sleep analysis in clinical practice and medical research. The results show that sleep can be reliably quantified in PD

and epilepsy patients, both for in-hospital and at-home monitoring. In the pediatric epilepsy population, automated sleep staging picks up on the sleep changes present in DRE, proving its usefulness. Furthermore, in AD patients, it is shown that automated sleep staging has potential beyond recognizing sleep changes, extending to therapeutic applications.

This chapter was made possible through collaborations with various departments from the University Hospitals Leuven and beyond. My special thanks go to L. Van den Bulcke, L. Micca, R. Proost, and J. Macea Ortiz.

8 | Wearable sleep recordings for Alzheimer's disease screening

This chapter is largely based on **E. R. M. Heremans***, A. Devulder*, P. Borzée, R. Vandenberghe, F.-L. De Winter, M. Vandenbulcke, M. Van Den Bossche, B. Buyse, D. Testelmans, W. Van Paesschen, and M. De Vos, "Home-based wearable sleep recording augmented by artificial intelligence for Alzheimer's disease screening: a method-development and proof-of-concept study," Submitted for review, 2024. (*: joint first authors)

ABSTRACT | Chapter 7 established that automated sleep scoring methods can characterize sleep in different clinical populations. The present chapter specifically focuses on AD, investigating the use of sleep recordings with a multimodal wearable device to screen for AD. To achieve this, we developed an AI model for AD detection based on sleep characteristics extracted through automated sleep staging on a wearable device. Measurements from PSG and wearable devices were simultaneously obtained from 67 older adults without cognitive symptoms and 35 AD patients. We calculated hypnogram features from the AI-predicted sleep stages, and physiological features quantifying spectral content in the measurements. An MLP was trained on these features to detect AD, and elastic net feature selection identified the most relevant features. The wearable AD detection model reached an area under the ROC-curve of 0.90 (0.76 for prodromal AD), with physiological features outperforming hypnogram features. We concluded that single-channel EEG and ACM from a wearable device can capture sufficient information for AD detection. Alternative approaches than hypnogram quantification, like the proposed physiological features, should be explored as discriminative markers for AD. Our approach shows promise for home-based, non-invasive AD screening.

8.1 Introduction

Around 50 million people worldwide live with dementia, of which AD is the main cause [167]. This public health problem has a huge economic impact [168]. Hence, there is an urgent need for sensitive markers to diagnose AD in the early disease stages to facilitate therapeutic interventions.

Sleep changes have been linked to AD pathology, even before cognitive symptoms emerge, endorsing its potential as an early noninvasive diagnostic feature for AD [169–172]. In a recent study from Ye et al. [173], a machine learning algorithm applied to PSG was able, to some extent, to distinguish patients with dementia or mild cognitive impairment from CIE individuals. As established in Chapter 1, PSG with manual five-stage sleep scoring would be impractical for large-scale screening purposes. Wearable devices might provide a more comfortable and feasible alternative, but require automatic data processing tools to manage the large data volumes. If we want to implement remote sleep measurements as a non-invasive screening tool for AD, the pivotal question is whether multimodal wearable devices can capture the necessary information. We need to determine if fully automated processing of wearable data can reliably detect individuals with AD in the older population.

To investigate if remote sleep monitoring with wearable devices can be used as a screening tool for AD on population level, the current study had three main objectives: firstly, to present an AI algorithm for automated sleep staging on AD patients based on gold-standard PSG and a wearable device measuring one-channel EEG and ACM; secondly, to design an AI-based algorithm that uses sleep features to discriminate AD patients in the prodromal and dementia disease stage from CIE individuals; and thirdly, to identify which sleep features are the most informative parameters for detecting AD.

RESEARCH IN CONTEXT

Evidence before the study | We searched Pubmed on April 16th 2024, from database inception, with the Mesh terms ("Wearable Electronic Devices" OR "Sleep" OR "Polysomnography") AND ("Dementia" OR "Alzheimer Disease")" AND ("Artificial Intelligence" OR "Machine Learning"). This revealed six results, all written in English. One editorial referred to work from Ye et al., who developed a machine learning algorithm to discriminate patients with cognitive impairment from CIE adults based on sleep measurements. However, this algorithm was developed on in hospital-acquired PSG and hence difficult to apply for large-scale screenings. Overall, our search did not generate publications covering the scope of our research, which was specifically focused on sleep measurements acquired with wearable devices to discriminate patients with AD from CIE participants.

Added value of this study | In this study, we developed an AI pipeline to distinguish AD from CIE based on sleep measurements that could be routinely obtained with a multimodal wearable device. The main technical novelty consisted in combining two AI algorithms for end-to-end, fully automated (early) AD screening based on signals from a wearable device, suitable for large-scale deployment. We trained an AI model for automated sleep staging and derived sleep stage-dependent features, which were in turn used to feed an AI classifier for AD detection. With this novel method, we made several discoveries with potential major clinical implications. First, we confirmed that sleep measurements can be used to screen for dementia, particularly AD, even in the prodromal disease stage. Second, we demonstrated that a wearable device, measuring single-channel EEG and ACM, can capture relevant information to identify key sleep features associated with AD. Third, we discovered that physiological EEG features, even in the case of imperfect sleep staging, are likely more sensitive discriminating features for AD, compared to macrostructural sleep parameters.

Implications of all the available evidence | With this proof-of-concept study, we demonstrated that sleep monitoring with wearable EEG devices, augmented with AI, holds great potential to be implemented in clinical practice as a feasible, non-invasive tool for AD screening. Our AD detection model, developed on a representative population, is ready to be extrapolated towards a large-scale home-based study aimed at screening for AD. Longitudinal sleep measurements with a wearable device will enable further exploration of these important sleep features, particularly towards screening for AD in the prodromal and even preclinical stage.

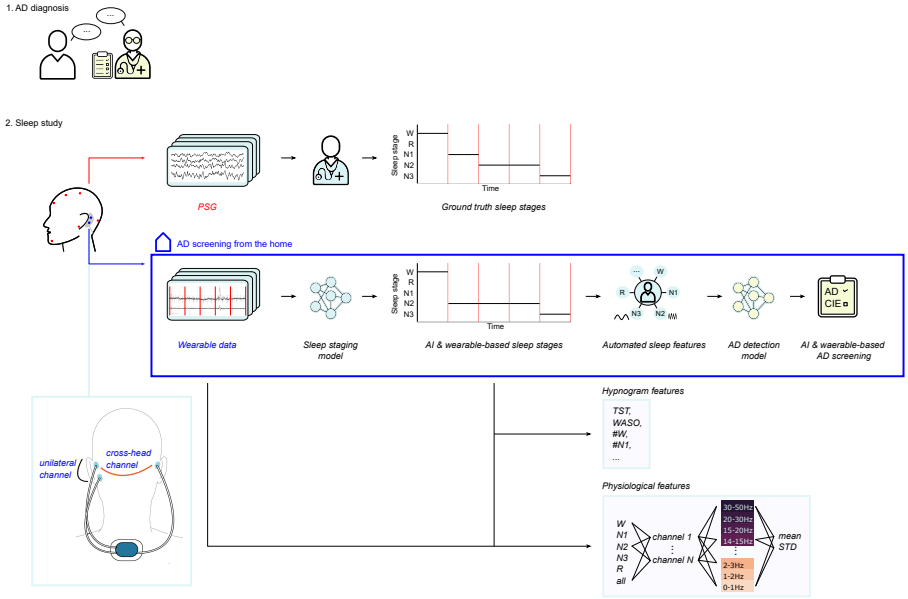


Figure 8.1: Overview of the methodology. Abbreviations: all, all sleep stages; abbreviations for hypnogram features are explained in Table 8.A.

8.2 Methods

8.2.1 Datasets and PSG scoring

For this method-development and proof-of-concept study (Figure 8.1), the dataset was pooled from the SDS and SDAD datasets (see Section 2.1.2), both encompassing overnight PSG simultaneously acquired with measurements from the SensorDot wearable device in older adults between 55 and 85 years old. Data were pooled from these two studies to increase the dataset size for training and testing our machine learning methods. A small portion of the SDS dataset was excluded in this chapter (details in Appendix 8.A).

The combined SDS dataset (Table 8.1) and SDAD dataset (Table 8.2) consisted of 147 eligible recordings. Thirty-five participants were diagnosed with either probable clinical (n = 7) or biomarker proven (n = 28) AD [74, 75] (Appendix 8.B). Thirty-four percent of patients with AD had an MMSE $\geq$ 27/30 and were further classified as prodromal AD [76]. The remaining 112 participants were free of subjective and/or objective cognitive symptoms and were considered

Table 8.1: Sample characteristics of the used portion of the Leuven – SensorDot Senior (SDS) dataset

	Train dataset	Test dataset
n	45	37
Age	65.96 ± 4.95	66.00 ± 5.10
% female	35.6	40.5
BMI*	33.44 ± 14.84	30.65 ± 6.58
AHI	33.38 ± 21.20	35.54 ± 25.66
ESS	8.31 ± 4.62	7.11 ± 4.15
% OSA	82.2	75.7

Abbreviations: ESS, Epworth Sleepiness Scale.
Data are reported as mean ± standard deviation unless otherwise indicated.
*BMI was not recorded for two patients of the test dataset.

Table 8.2: Sample characteristics of the Leuven – SensorDot AD (SDAD) dataset

	non-AD	AD	
		Total	Prodromal AD*
n	30	35	12
Age	70.63 ± 7.10	74.63 ± 6.52	74.5 ± 6.29
% female	63.3	45.7	33.3
BMI	24.69 ± 3.84	23.52 ± 3.25	24.28 ± 3.36
AHI	8.07 ± 6.71	16.81 ± 13.54	21.2 ± 15.97
ESS	4.40 ± 3.04	4.74 ± 3.56	4.17 ± 3.51
MMSE	28.87 ± 0.88	20.00 ± 7.33	27.92 ± 1.04
PSQI	4.77 ± 3.24	3.86 ± 3.29	4.42 ± 3.59
% OSA	20.0	51.3	75.0

Abbreviations: PSQI, Pittsburg sleep quality index.
Data are reported as mean ± standard deviation unless otherwise indicated.
*12 of the AD patients had an MMSE score $\geq$ 27/30 and were labeled as being in the prodromal AD stage.

CIE. The combined datasets were divided in a training and test dataset. The training dataset (n = 45, Table 8.1) was only used for training a sleep staging model, and the test dataset (n = 102, Table 8.3) was used for all subsequent analyses, including cross-validation of the AD detection models.

Table 8.3: Sample characteristics of the test dataset

	CIE - no OSA	CIE - OSA	AD - no OSA	AD - OSA
n	33	34	17	18
Source	9(SDS) + 24(SDAD)	28(SDS) + 6(SDAD)	17(SDAD)	18(SDAD)
Age	67.79 ± 7.38	68.45 ± 5.49	73.00 ± 7.19	76.17 ± 5.39
% female	66.7	35.3	58.8	33.3
BMI	25.75 ± 4.28 (32*)	29.98 ± 7.09 (33*)	23.45 ± 2.54	23.59 ± 3.79
AHI	6.11 ± 4.41	39.87 ± 23.31	6.28 ± 4.60	26.76 ± 11.52
ESS	5.36 ± 3.72	6.41 ± 4.06	5.12 ± 3.72	4.39 ± 3.37
MMSE	28.88 ± 0.88 (24*)	28.83 ± 0.90 (6*)	17.76 ± 6.56	22.11 ± 7.38

Abbreviations: SDAD, Leuven - SensorDot Alzheimer's disease sleep study; SDS, Leuven - SensorDot Senior sleep study.
Data are reported as mean ± standard deviation unless otherwise indicated.
*When a clinical parameter was not recorded in all the samples, the number in between brackets indicates for how many samples it was recorded.

8.2.2 Sleep staging model

SeqSleepNet [46] (see Section 2.2), was trained to perform automated sleep staging on two channels of the wearable data: the cross-head EEG and ACM (Figure 8.1). It was trained on the training dataset of 45 recordings and then deployed to score the entire test dataset of 102 recordings for subsequent sleep feature calculation and AD detection.

8.2.3 Feature calculation

Based on the sleep stages and the measured signals, we computed two sets of sleep features: hypnogram features and physiological features (Figure 8.1). Thirty-four hypnogram features were calculated from the hypnograms, representing macrostructural sleep parameters (Appendix 8.A). Physiological features were spectral features of the measured signals, including EEG and ACM for the wearable. They represented averages of the frequency content per channel, sleep stage, and frequency bin of the raw signals' spectrogram (Appendix 8.A).

8.2.4 Detection of patients with AD

We employed an MLP to assess the diagnostic potential of both the hypnogram and physiological features in discriminating AD from CIE (Appendix 8.A). We aimed to evaluate AD detection performance using wearable data scored by AI in comparison to two baselines: PSG scored by AI and by a certified scorer. Therefore, both the hypnogram and physiological features were calculated with

three approaches: using the PSG with manual (ground truth) sleep scoring, using the PSG with AI scoring, and using the wearable data with AI scoring. We evaluated the ability to discriminate all AD patients from CIE adults, and additionally, to distinguish patients in the prodromal AD stage from the CIE. Both the AD and prodromal AD classifier were evaluated using ten repeats of leave-one-out cross-validation. The classifiers were primarily evaluated with receiver operating characteristic (ROC) curves and area under the curve (AUC) as aggregate performance measure (Appendix 8.A).

To gain deeper insight into the sleep features' relation to AD and explore their clinical significance, we performed two distinct analyses: a correlation analysis with Mini-Mental State Examination (MMSE) score [76] and an AI-based feature importance analysis. Firstly, the sleep features' clinical relevance was assessed by computing the Pearson correlation with MMSE for all hypnogram features as well as for a small subset of physiological features (Appendix 8.A). Secondly, we trained the MLP with elastic net feature selection (Appendix 8.A), identifying the most informative features for AD detection with ten leave-one-out cross-validation runs and reporting the most frequently selected ones.

8.3 Results

8.3.1 Sleep staging results

The overall five-stage sleep staging performance of the sleep staging model was 65.5% on the multimodal wearable device, with a κ of 0.498, indicating moderate agreement. This was an absolute 11.4% lower than the model performing sleep staging on PSG, and it could be improved by 4.8% by further fine-tuning of the model (Appendix 8.B).

8.3.2 Detection of patients with AD

Figure 8.2 shows the ROC curves of the AD classification results. The physiological features (Figure 8.2d, e and f) outperformed the hypnogram features (Figure 8.2a, b, and c) by far in every scenario, with differences in AUC of 0.26 to 0.37.

Despite imperfect sleep staging, the classification performance with AI-estimated physiological features, both PSG- and wearable-based, was almost on par with the ground truth features. Strikingly, based on a wearable with only one channel of EEG and ACM, scored by AI, our MLP was able to detect patients with

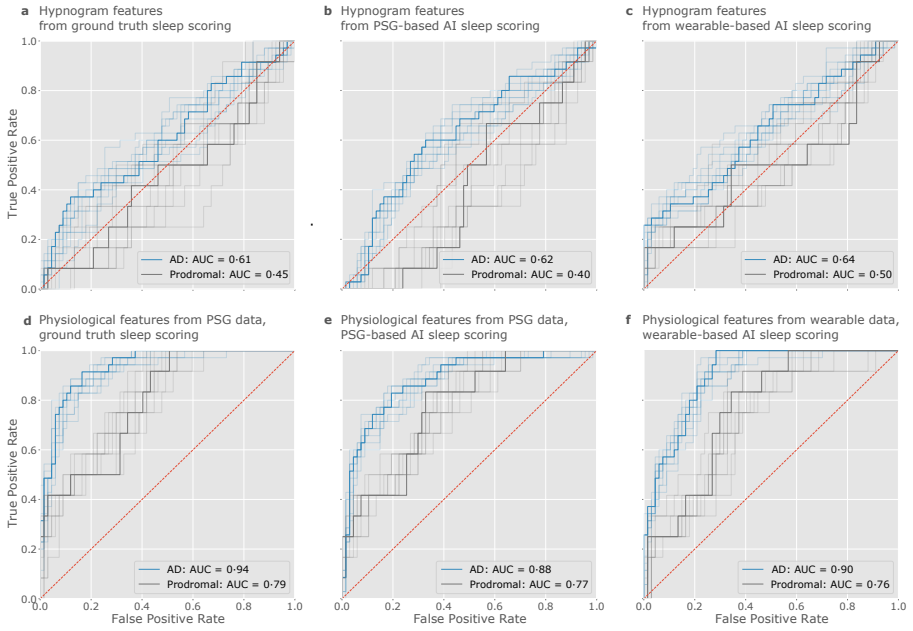


Figure 8.2: The performance in detecting patients with AD and the subset with prodromal AD. Transparent lines represent the ROC curves of all ten cross-validation runs for each prediction task, while the non-transparent lines represent the ROC curves obtained using the mean predictions with these ten repeats. Their AUC is reported in the legend. As the prodromal AD patients are a subgroup of AD, the curves labeled with 'AD' show the performance in discriminating CIE patients from the whole AD cohort, consisting of patients in the prodromal and dementia disease stage. The curves labeled with 'prodromal' show the performance in discriminating the prodromal AD patients from the CIE group. The upper row shows performances using hypnogram features, calculated based on (a) manual PSG labels (ground truth), (b) AI-labeled PSG, (c) AI-labeled wearable data. The lower row shows performances using physiological features, calculated using both the raw data and sleep stage labels: (d) raw PSG data with manual labels, (e) raw PSG data with AI-based labels, and (f), wearable data with AI-based labels. The ROC curves show that the physiological features by far outperform the hypnogram features. For the physiological features, the AI-estimated features (based on both PSG and wearable) are almost on par with the ground truth features.

AD with an AUC of 0.90, and specifically prodromal AD patients with AUC 0.76. The performance as measured by AUC was similar for the MLP based on AI-scored PSG, although it benefited from four extra measurements. Training a classifier on the combined set of the hypnogram and physiological features did not improve the performance. As OSA is associated with sleep disruption, we

also compared the performance in detecting AD patients with and without OSA and found similar performances for both groups (Appendix 8.B). Additionally, we investigated the effect of age and dataset on the performance. Different performance metrics were also computed (Appendix 8.B).

8.3.3 Insights into the hypnogram and physiological features

Figure 8.3 shows the correlations between the MMSE score and the sleep features. Based on ground truth scoring, 42.9% of the hypnogram features and 32.2% of the physiological features showed significant correlations with MMSE. The wearable-based AI model matched the significant correlations of the hypnogram and physiological features in 64.7% and 82.2% of the cases, respectively. Utilizing PSG data, the concordance between AI-based features and ground truth features was even higher than for the wearable data (Appendix 8.B).

We used feature selection to examine which features were the most informative. The assessment for the hypnogram features was performed based on the ground truth features, the PSG-based AI-derived features, and the wearable-based AI-derived features. From the ground truth hypnogram features, the four most selected were mean time between consecutive REM bouts, number of transitions to N3 divided by TST, WASO, and mean N3 bout duration (N3BoutDuration). Among the PSG-based AI-estimated hypnogram features, the latter two and the amount of N1 and N2 were the most selected. For the wearable-based AI-estimated hypnogram features, the amount of N2 and REM were the most informative, along with the number of transitions to REM and N3BoutDuration. We concluded from the hypnogram features that indicators of AD in the sleep architecture were mainly the distribution of N3 and REM (including number of transitions and bout durations), as well as WASO, and amount of light sleep.

The informativeness of the physiological features was evaluated in the same manner. Figure 8.4 displays features that were selected more than 85% of times by the AD detection model based on the wearable data (Figure 8.4a) and PSG data (Figure 8.4b). The wearable-based feature selection indicated that important EEG features included 9 - 11 Hz in Wake, 4 - 5 Hz in N1, 9 - 10 Hz in N2, and 0 - 1 Hz in REM. Derived from PSG, the most important EEG features included 4 – 6 Hz in Wake, 1 – 2 Hz and 7 – 8 Hz in both N1 and N2, 0 – 1 Hz and 11 – 14 Hz in N3, and 0 – 1 Hz in EOG in REM. From the wearable's ACM and PSG's EMG, features were selected in Wake, N1, and the whole recording. From both the wearable and the PSG, we can conclude that the alpha and theta EEG band in wakefulness and light sleep are helpful to detect AD, as well as slow activity during REM. Additionally, slow delta activity (0 – 2 Hz) in

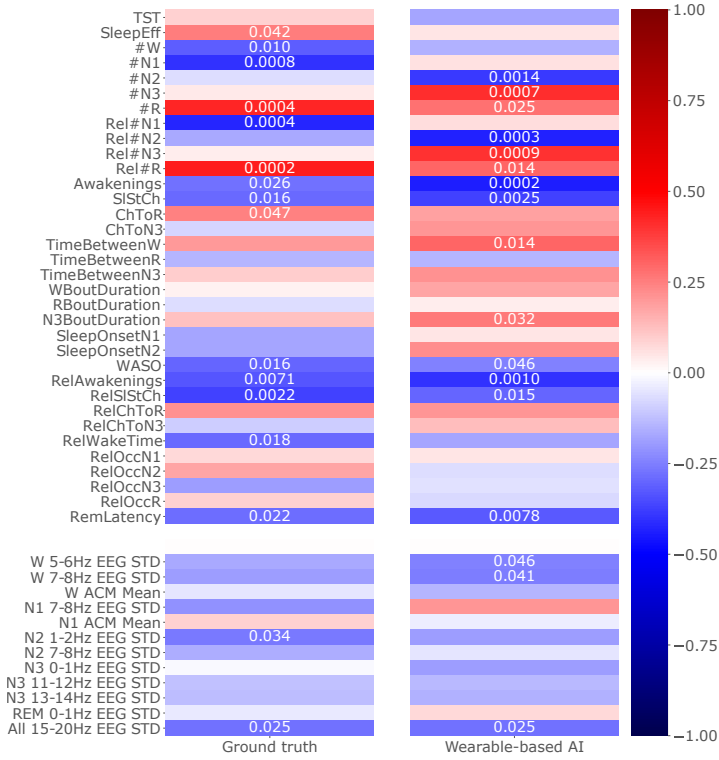


Figure 8.3: Correlation between MMSE and sleep features based on wearable data. The hypnogram features are shown in the upper part of the figure, and a few selected physiological features are shown in the lower part. The first column shows the ground truth sleep features based on the wearable data, but scored with the ground truth manual PSG scoring, and the second column shows sleep features based on the wearable data scored by AI. Pearson's correlation test was performed, with the subjects as samples. The MMSE score was not recorded in the patients from the senior sleep dataset, so the correlations are only computed for the 65 patients of the SDAD dataset (see Table 8.2). Correlations are shown through the colors, with p values below 0.05 reported as numbers to indicate the significance of the correlations. The wearable-based AI features and ground truth features agree on most significant correlations. Abbreviations: All, all sleep stages; STD, standard deviation; abbreviations for hypnogram features are explained in Table 8.A.

non-REM sleep also holds discriminative power, although this was only picked up on in PSG. A comparison between features selected using ground truth and AI scoring highlighted imperfections in sleep staging, but confirmed that key features remained identifiable and valuable for AD detection (Appendix 8.B).

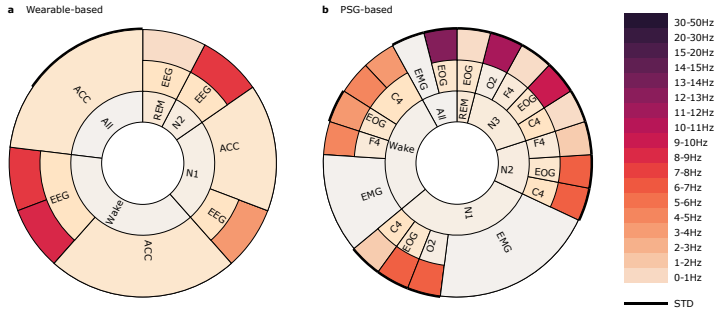


Figure 8.4: Most selected wearable-based and PSG-based physiological features for the AD detection model. The features that were selected more than 85% of the time by the (a) wearable-based and (b) PSG-based AD detection model. In both cases, the features were calculated using ground truth sleep scoring (for the features calculated using AI scoring, see Appendix 8.B). Features are visualized using three rings showing to which (1) sleep stage, (2) channel, and (3) frequency range each feature corresponds. The frequency range is only shown for the EEG and EOG channels due to the limited interpretability of the EMG and ACM frequencies. The discrete color bar shows the interpretation of the frequency range. The surrounding black ring shows which features were computed using STD, with the remaining calculated using the mean. The width of each segment indicates how many features of that type were selected more than 85% of the times. This width should not be mistaken for overall importance of a channel or sleep stage: information may be more spread out over multiple frequency bands for some channels or sleep stages than for others. Rather, this figure shows which channels are important in which sleep stages, and which frequencies were decisive in those channels. Abbreviations: All, features computed over the whole recording; STD, standard deviation.

8.4 Discussion

In this study, we investigated the viability of home-based multimodal automatic sleep monitoring in the elderly population, specifically towards screening for AD. For the first time, we developed a screening tool for AD based on a multimodal wearable device, measuring single-channel EEG and ACM, by utilizing a fully automatic AI pipeline. First, automatic sleep scoring was conducted to extract relevant sleep features, which were then leveraged for AI-based AD detection. We validated our pipeline on both PSG and wearable device, showcasing its versatility across hospital and home environments. To gain insight into which sleep features were most predictive for AD, we investigated which sleep features correlated most with MMSE, and which sleep features our AI model selected the most for AD detection.

8.4.1 AD detection performance

A unique feature of our study is that we distinguished patients with AD from older individuals without cognitive complaints based on measurements from a wearable device. With the objective of large-scale AD screening in mind, we designed a fully automated AI pipeline for this purpose. With an AUC of 0.90, our wearable-based AI model demonstrated excellent performance for AD detection (see Figure 8.2f). This was especially remarkable as we included sleep apnea as a potential confounder. The results were on par with the performance in detecting AD with the same AI approach based on PSG measurements (see Figure 8.2e). This shows that single-channel EEG with ACM may be sufficient for AD detection, and that perfect sleep staging is not needed to achieve a useful screening tool. When using physiological features, sleep staging primarily served to delineate bins for averaging the spectral content of measured signals. Hence, the spectral content itself was more important than how it was categorized into sleep stages. Conversely, for hypnogram features, the accuracy of sleep staging played a more pivotal role, as was expected. This was evidenced by the larger performance difference between features derived from manual and automatic scoring.

Our performances were higher compared to those of the dementia detector from Ye et al. [173], even though they relied on manually scored PSG. First, the difference might be explained by different features used: they only used means to average the EEG content, whereas we found variability of the features to be more informative. Second, we specifically aimed to detect AD, while their study focused on dementia of all types. Although their model was developed on a larger PSG sample, they included diagnostic and CPAP titration PSGs from the same individuals and there was a greater class imbalance, which may have influenced their results.

Another non-invasive and cost-effective method for screening AD is through clinical data and questionnaires. A study involving adults with memory complaints demonstrated that using extensive clinical data and questionnaires to test cognitive abilities, individuals who would develop AD within a few years could be detected with an AUC of 0.85-0.87 [174]. As the difference in population and methodology doesn't allow direct comparison of these results with the ones in this chapter, we performed our own small experiment (see Section 8.B: Figure 8.E) in the SDAD dataset. This further confirmed that clinical questionnaires and data hold substantial discriminatory power and significantly enhanced the performance of our sleep-based AD detector. From this experiment, we conclude that physiological measurements of sleep, clinical data, and questionnaires could potentially be combined to achieve better discrimination in AD screenings.

When training our classifier model to specifically detect the subgroup of patients in the prodromal disease stage, the AUC dropped to 0.79 for the PSG-based model and to 0.76 for the wearable-based model. A drop in performance was expected for two reasons. First, sleep and EEG alterations become more pronounced as AD progresses [175,176], as evidenced by the correlations with MMSE scores (Figure 8.3) [177]. Second, the amount of prodromal AD patients was low (n = 12). Training the model on a larger sample of prodromal AD patients should increase the performance.

8.4.2 Sleep features related to AD

The most informative hypnogram features to distinguish AD from CIE were very similar between the ground truth and wearable-based AI scoring. Features related to N3, REM and WASO were most relevant. For the wearable-based AI model, an increased N2 was selected as well, likely due to the algorithm confusing REM for N2 in AD patients. Overall, the selected hypnogram features reflect the macrostructural sleep changes observed in AD, consistent with literature [73,175,177].

The most valuable physiological features selected for the PSG-based model, using ground truth sleep scoring, ensued from all sleep stages, with EEG frequencies around 0 - 2 Hz, 4 - 6 Hz, 7 - 8 Hz and 11 - 14 Hz being most selected. For the wearable-based model using ground truth scoring, the most informative physiological features derived from the EEG were mainly found in wakefulness, light sleep and REM. Overall, these features are congruent with the most important features from the model of Ye et al. [173] and reflect an increased slowing of the EEG, consistent with the microstructural sleep alterations observed in AD [176,178,179].

Interestingly, the physiological features outperformed the hypnogram features in both models. Combining the two feature sets did not help the models to achieve a higher accuracy, and even the hypnogram features based on manual scoring demonstrated inferior performance compared to the physiological features based on the wearable with AI scoring. For prodromal AD detection, the difference was even more pronounced, with near-chance-level accuracies achieved based on hypnogram features. Microstructural sleep changes without obvious macrostructural sleep changes have been observed in the prodromal disease stage [176,178]. This might indicate that microstructural sleep alterations, as reflected in the physiological EEG features, could be considered as more promising biomarkers for detecting patients with AD compared to macrostructural changes, especially in the early disease stages.

8.4.3 Automatic sleep scoring

The PSG- and wearable-based AI model for automatic sleep scoring reached an accuracy of 76.9% and 65.5%, respectively, compared to gold standard manual PSG scoring. This difference in accuracy was not surprising as we went from five channels to two channels, omitting EOG, EMG and 2 channels of EEG. The two sleep stages that were mostly responsible for this drop in performance were N1 and REM. In retrospect, recording a second channel at F7 or F8 to detect eye movements may have improved N1 and REM sleep detection. Although REM was still scored well in CIE patients, the performance drop was large in patients with AD, where REM was frequently misclassified as N2. This could be attributed to the REM slowing, which is often present in AD [176,178,179]. The 65.5% accuracy for five-stage sleep scoring was slightly lower compared to previous studies on one- or two-channel EEG recordings [29], but those studies were conducted in smaller samples consisting of healthy younger individuals. The presence of OSA and AD in our dataset, alongside a more advanced age, are associated with a more disrupted sleep [147,175,177], which is more difficult to score. This was shown by the 5% increase in accuracy when evaluating our model on the CIE subgroup without OSA Appendix 8.B. We observed that fine-tuning to our population also improved the performance significantly. We conclude that sleep staging performance is influenced by numerous factors related to patient demographics and recording device. However, sleep staging should not be the sole focus of automated sleep analysis, as we showed that the resulting sleep parameters can effectively detect AD with comparable accuracy, even with the sleep staging accuracy varying from 65.5% to 76.9%.

8.4.4 Implications for future research

Based on our results, this study holds potential implications for public health and avenues for future research. Firstly, consistent with previous studies [29], we demonstrated that one-channel EEG devices can present a reliable, comfortable alternative compared to the cumbersome gold-standard PSG, and, consequently, could be applied for longitudinal home-based sleep monitoring. Secondly, now that we have established that AI algorithms can detect prodromal AD based on home-based wearable sleep measurements, it would be useful to extend this research to large cohorts in the preclinical AD stage. If successful, we can apply sleep measurements as a screening tool to detect patients in the preclinical AD stage, which in turn might offer opportunities for therapeutic trials. Thirdly, since the physiological features were more sensitive than the hypnogram features, future sleep studies should deploy AI approaches that identify microstructural sleep changes. Fourthly, sleep disturbances and epilepsy have an increased

prevalence in AD. Decreased REM, a potential epilepsy biomarker, could be monitored in AD using a wearable device [73, 180].

8.4.5 Limitations and strenghts

Our study sample had some limitations. The criteria used to define CIE differed for both datasets used. For the SDS dataset, we retrospectively searched the electronic health records for reported cognitive complaints, excluding participants with such complaints. In a subsample, no information about cognitive function was available. For both datasets, no neuroimaging or CSF biomarkers were available for the participants considered CIE, so preclinical AD could not be ruled out. In addition, AD biomarkers were not available for twelve patients of the AD group. Therefore, the presence of another underlying pathology can't be excluded. Despite the limitations inherent to this study sample, a main strength of this sample was the fact that it consisted exclusively of older participants with a balanced ratio of patients with OSA and a large proportion of patients with AD. By incorporating challenges posed by OSA and older age, known risk factors for AD and obstacles for automated sleep staging algorithms, we constructed a sample reflecting real-world complexities. While larger sample sizes are warranted for future validation of our AD detection method, our algorithm development on this representative population enables reliable extrapolation to a large-scale home-based study. Another strength of this study was the fully automated nature of the proposed method and its suitability for home-based monitoring. Whereas previous studies have relied on full-scalp EEG or PSG with manual sleep scoring, our automated, wearable approach represents a significant advancement towards integrating sleep as a screening tool in future clinical practice.

8.5 Conclusion

We demonstrated that AI algorithms allow to discriminate patients with AD from CIE individuals based on sleep measurements, taking OSA into consideration. Physiological features, specifically those related to EEG slowing in wakefulness, light sleep, and REM, and a decrease in slow wave activity during N3, are the most informative diagnostic features for AD. In addition, with an AI model trained on a wearable device, we demonstrated that one-channel wearable EEG and ACM contain sufficient information and that perfect five-stage sleep scoring is not required for AD detection. Further refinement of our proposed method in larger prodromal AD cohorts, along with extension to the preclinical AD population, offer promising avenues for early AD screening in older adults.

Appendices

8.A Supplementary methods

SDS dataset details

For a full description of the SDS dataset, see Section 2.1.2. A portion of the SDS dataset was used as a training set (to pre-train our sleep staging algorithm). The rest was pooled with the SDAD dataset as a test set, for evaluating the sleep staging performance and for training and testing an AD detection model. In the training set, cross-head EEG recordings from the wearable device were not available (see Section 2.1.2), so behind-the-ear EEG data from standard EEG equipment were used for those subjects.

Clinical data extraction from medical records revealed that 18 participants of the SDS dataset reported subjective cognitive complaints, 8 of whom were part of the test dataset. Subsequently, these eight individuals were excluded from the test set to eliminate ambiguity concerning AD. Of the 72 participants without subjective cognitive complaints, sixteen had no subjective cognitive complaints, and for the other 56, no information about subjective or objective cognitive functioning was available in the medical record.

Sleep staging model: architecture and training

SeqSleepNet was trained for sleep staging on both the PSG data and wearable data of our datasets. The PSG-based model used five channels of the PSG data: the C4-P9, F4-P9, O2-P9, chin EMG1-2, and EOGleft-EOGright. For the wearable-based model, we used 2 channels: the cross-head EEG channel and the ACM channel. Both for sleep staging on the PSG and wearable data, the model was pre-trained on the 45 recordings of the training set from the SDS dataset for 10 training epochs. The pre-trained model was directly applied to score the entire test dataset for subsequent feature calculation and AD detection (Appendix 8.A). Additionally, to assess the maximum sleep staging performance across the patient subgroups, we fine-tuned the model on the test dataset, conducting ten-fold cross-validation. We then personalized the model to each individual recording in the test dataset using unsupervised adversarial domain adaptation, following the approach from Chapter 4. Hence, the personalization step did not use any labels from these test patients. In conclusion, the sleep staging results (Appendix 8.B) were based on the sleep scoring by the fine-tuned, personalized models. The rest of the results used the pre-trained models' scorings.

Hypnogram features calculation

Table 8.A shows all 34 hypnogram features that were calculated. Most features have an obvious definition, except the average timing of different sleep stages ('RelOccX' with X being the sleep stage), which was calculated as:

$$RelOccX = \frac{mean(where(y == X)) - onset_{sleep}}{end_{sleep} - onset_{sleep}} \tag{8.1}$$

with y being the annotated sleep stages, and the function $where$ returning the times when y is equal to the sleep stage X.

Physiological features calculation

Physiological features were spectral features of the EEG and other measured signals (EMG and EOG in the case of PSG, and ACM in case of the wearable). They were derived from the STFTs of the raw signals, which were computed with 129 frequency bands (representing frequencies between 0 and 50 Hz). To calculate these features, we generated individual time series for each frequency component within every channel for each subject or recording. In the feature space, we reduced the amount of frequency bins to 18, by averaging features within 1 Hz intervals from 0 to 15 Hz, and within the larger intervals of 15 - 20 Hz, 20 - 30 Hz and 30 - 50 Hz. Temporally, the time series were aggregated per sleep stage (W, N1, N2, N3, R) and for the total recording, using the mean and standard deviation as aggregate measures. For the PSG, the total amount of physiological features was 1080: the number of frequency bands (18) $\times$ the number of channels (5) $\times$ the number of sleep stages +1 (6) $\times$ 2, because both the mean and standard deviation were computed for each time series. For the wearable, the number of physiological features was 432: 18 $\times$ 2 $\times$ 6 $\times$ 2.

AD detection model: architecture and training

We trained a classifier to discriminate AD from CIE. For both sets of features (physiological and hypnogram), classification was performed with a MLP with a hidden layer of size 20. An MLP was selected because of its strong ability to model complex, nonlinear relationships, and its capacity to generalize to unseen data, especially in high-dimensional datasets. The rest of the parameters were the standard parameters in the Scikit-learn library [181]. The optimization was performed for 200 iterations, the batch size was the full dataset, the ReLU activation function was used, and the Adam optimizer was used with a learning rate of 10^{-3}. The presence of class imbalance necessitated the consideration of

Table 8.A: Hypnogram features and their meaning.

Abbreviation	Meaning
TST	Total sleep time (hours)
SleepEff	Sleep efficiency: TST/ time in bed
#W	Amount of time in Wake (hours)
#N1	Amount of time in N1 (hours)
#N2	Amount of time in N2 (hours)
#N3	Amount of time in N3 (hours)
#R	Amount of time in REM (hours)
Rel#N1	Relative time in N1: #N1 /TST
Rel#N2	Relative time in N2: #N2 /TST
Rel#N3	Relative time in N3: #N3 /TST
Rel#R	Relative time in REM: #R /TST
Awakenings	Total number of awakenings, i.e. transitions to Wake
SlStCh	Total number of sleep stage changes
ChToR	Total number of transitions to REM
ChToN3	Total number of transitions to N3
TimeBetweenW	Average time between Wake bouts (hours)
TimeBetweenR	Average time between REM bouts (hours)
TimeBetweenN3	Average time between N3 bouts (hours)
WBoutDuration	Average Wake bout duration (hours)
RBoutDuration	Average REM bout duration (hours)
N3BoutDuration	Average N3 bout duration (hours)
SleepOnsetN1	Sleep onset as time until first N1 epoch (hours)
SleepOnsetN2	Sleep onset as time until first N2 epoch (hours)
WASO	Wake after sleep onset: Wake time from sleep onset to final awakening
RelAwakenings	Relative number of awakenings: awakenings/TST (/hour)
RelSlStCh	Relative number of sleep stage changes: SlStCh/TST (/hour)
RelChToR	Relative number of transitions to REM: ChToR/TST (/hour)
RelChToN3	Relative number of transitions to N3: ChToN3/TST (/hour)
RelWakeTime	Relative time spent awake after sleep onset: WASO/TST
RelOccN1	Average timing of N1 relative to sleep onset and final awakening
RelOccN2	Average timing of N2 relative to sleep onset and final awakening
RelOccN3	Average timing of N3 relative to sleep onset and final awakening
RelOccR	Average timing of REM relative to sleep onset and final awakening
RemLatency	Time until first REM epoch after sleep onset (hours)

multiple performance metrics. Performance was primarily assessed using ROC-curves and their AUC, with additional computation of accuracy, κ, precision, and recall.

Pearson correlation details

We investigated Pearson correlations of sleep features with MMSE to investigate the features' clinical relevance and to verify whether potential correlations were preserved with AI scoring. The MMSE was only available for the 65 patients from the SDAD dataset (and not for the 37 patients from the SDS dataset), so the Pearson correlation was only computed for 65 patients, with the patients being samples in the test. p values under 0.05 were reported.

The correlation test was performed for sleep features based on wearable data with ground truth (PSG-based) scoring and with wearable-based AI scoring (results in Figure 8.3) and for sleep features based on PSG data with PSG-based ground truth scoring and with PSG-based AI scoring (results in Appendix 8.B). The ground truth hypnogram features in both the wearable and PSG results are the same, as they only depend on the scoring (the ground truth scoring being the manual PSG scoring), whereas the ground truth physiological features are different for the wearable and PSG results, because they are based on both the scoring and the data.

Feature selection details

Elastic net feature selection is an embedded feature selection method, which means that the model learns the importance of features while it gets trained. This feature selection method was chosen because of the highly correlated nature of the feature sets. We performed feature selection for both the hypnogram and physiological features, with a standard L1-ratio of 0.5. The regularization strength alpha was determined by conducting grid search on each cross-validation training set, to determine the best alpha value in the range of 10^{-5} to 10. Ten leave-one-out cross-validation runs led to a total of 10×110 iterations of feature selection for each feature set. Only features selected more than 85% of times were reported. Feature selection was solely used to determine which features were most informative in this study. As the AD classification performance did not consistently benefit from using feature selection, we reported the AD detection results without feature selection.

For the hypnogram features, feature selection was performed on three sets of features: (1) the ground truth features, (2) the features from AI scoring on

wearable data, and (3) the features from AI scoring on PSG data. The results for all three feature sets were reported in the main text. For the physiological features, four sets of features were considered: (1) the features based on PSG data and ground truth scoring, (2) the features based on PSG data and PSG-based AI scoring, (3) the features based on wearable data and ground truth (PSG-based) scoring, and (4) the features based on wearable data and wearable-based AI scoring. The results for all four feature sets were reported in Appendix 8.B.

Statistics implementation

All modeling, analysis, and statistics were performed in Python (version 3.7.9). The SciPy toolbox12, version 1.6.2, was used as to compute the t-tests and Pearson correlations. The Scikit-learn library [181] (version 0.23.2) was used for training the AD detection MLP, and TensorFlow13 (version 1.15.0) was used for training the deep learning sleep staging model.

8.B Supplementary results

Biomarkers of the AD patients

Table 8.B: Biomarkers for the AD patients of the Leuven - SensorDot Alzheimer's disease (SDAD) dataset.

Nb	Disease stage	Amyloïd-PET	CSF Biomarkers	18F-FDG-PET	Brain MRI (or CT)
1	Dementia	NA	t-tau = 365 p-tau 181 = 52.7 Aβ 1-42 = 501 (a)	Hypometabolism temporoparietal, precuneus and frontal	Normal for age
2	Dementia	NA	NA	NA	Generalized atrophy. Mild mesial temporal and parietal atrophy
3	Prodromal	Positive	Aβ 1-42 = 442 Ratio 1-42/1-40 = 0.087 p-tau = 90.4 t-tau = 592 (b)	Mild hypometabolism temporal and parietal bilateral	Normal for age
4	Dementia	NA	NA	NA	Generalized atrophy. Prominent mesiotemporal atrophy.
5	Dementia	NA	NA	NA	Generalized atrophy, frontotemporal predominance. Prominent mesial temporal atrophy R >L
6	Prodromal	NA	Ratio 1-42/1-40 = 0.076 Aβ 1-42 = 410 t-tau = 275 (c)	No hypometabolism	Mild generalized atrophy. Moderate vascular white matter lesions.

Table 8.B: Clinical biomarkers for the AD patients of the Leuven - SensorDot Alzheimer's disease (SDAD) dataset - continued

Nb	Disease stage	Amyloïd-PET	CSF Biomarkers	18F-FDG-PET	Brain MRI (or CT)
7	Dementia	NA	t-tau = 581 p-tau = 80.9 Aβ 1-42 = 695 (a)	Hypometabolism frontal R, bilateral parietal and cuneus R >L	Mild generalized atrophy. Mild vascular white matter lesions. FLAIR hyperintensity frontal right, parietal left and insula (old ischemic lesions).
8	Dementia	NA	t-tau = 770 p-tau = 107.8 Aβ 1-42 = 332 (a)	Hypometabolism temporoparietal R >L, precuneus and frontal L	Generalized atrophy, parietal and temporal predominance. Prominent mesial temporal atrophy R >L
9	Prodromal	NA	Aβ 1-42 = 452 Aβ 1-40 = 6765 Ratio 1-42/1-40 = 0.067 t-tau = 425 p-tau = 58 (c)	NA	Frontoparietal atrophy. Mild mesial temporal atrophy
10	Dementia	NA	NA	NA	Prominent temporoparietal atrophy. Prominent mesial temporal atrophy L >R
11	Dementia	Positive	NA	NA	Normal for age

Table 8.B: Clinical biomarkers for the AD patients of the Leuven - SensorDot Alzheimer's disease (SDAD) dataset - continued

Nb	Disease stage	Amyloïd-PET	CSF Biomarkers	18F-FDG-PET	Brain MRI (or CT)
12	Dementia	NA	NA	NA	Generalized atrophy, parietal predominance. Prominent mesial temporal atrophy L >R. Mild to moderate white matter lesions.
13	Prodromal	NA	NA	NA	Generalized atrophy. Prominent mesial temporal atrophy R >L. Prominent white matter lesions.
14	Prodromal	Positive	NA	Hypometabolism temporoparietal L >R, frontal L, precuneus L >R	Parietal atrophy. Mild white matter lesions.
15	Dementia	NA	NA	Mild hypometabolism frontal L	Mild generalized atrophy. Mild to moderate white matter lesions. Old lacunar infarct thalamus L
16	Dementia	NA	NA	Hypometabolism bilateral parietal and mild temporoparietal L	Generalized atrophy, temporoparietal predominance.
17	Dementia	NA	NA	Hypometabolism parietotemporal L >R, frontal L and precuneus bilateral	Mild generalized atrophy. Symmetric temporal lobe atrophy.

Table 8.B: Clinical biomarkers for the AD patients of the Leuven - SensorDot Alzheimer's disease (SDAD) dataset - continued

Nb	Disease stage	Amyloïd-PET	CSF Biomarkers	18F-FDG-PET	Brain MRI (or CT)
18	Prodromal	NA	Aβ 1-42 = 361 Aβ 1-40 = 7026 Ratio 1-42/1-40 = 0.051 t-tau = 715 p-tau = 110 (c)	NA	Generalized atrophy. Prominent mesial temporal atrophy. Moderate vascular white matter lesions. Gliosis frontoparietal right.
19	Dementia	NA	NA	NA	Prominent mesial temporal atrophy L >R. Periventricular leukoencephalopathy
20	Dementia	NA	Aβ 1-42 = 388 Aβ 1-40 = 7588 t-tau = 362 (c)	NA	Generalized atrophy, frontoparietal predominance. Prominent mesial temporal atrophy. Mild to moderate white matter lesions.
21	Prodromal	NA	NA	NA	CT: Mild parietal atrophy.
22	Dementia	NA	NA	Hypometabolism temporoparietal R >L, precuneus and posterior cingulate bilateral	Generalized atrophy. Moderate mesial temporal atrophy. Mild tot moderate white matter lesions.

Table 8.B: Clinical biomarkers for the AD patients of the Leuven - SensorDot Alzheimer's disease (SDAD) dataset - continued

Nb	Disease stage	Amyloïd-PET	CSF Biomarkers	18F-FDG-PET	Brain MRI (or CT)
23	Dementia	NA	NA	NA	Generalized atrophy. Moderate mesial temporal atrophy. Mild to moderate periventricular white matter lesions.
24	Dementia	NA	Aβ 1-42 = 309 Aβ 1-40 = 7663 Ratio 1-42/1-40 = 0.040 t-tau = 1034 (c)	NA	Generalized atrophy, frontoparietal predominance. Prominent mesial temporal atrophy. Mild to moderate white matter lesions.
25	Dementia	NA	NA	NA	Generalized atrophy. Prominent mesial temporal atrophy L >R. Mild to moderate white matter lesions.
26	Prodromal	NA	Aβ 1-42 = 667 Aβ 1-40 = 8838 t-tau = 674 (c)	Hypometabolism parietotemporal bilateral, frontal bilateral, precuneus and posterior cingulate.	Mild generalized atrophy. Prominent parietal atrophy. Mild mesial temporal atrophy L >R. Mild vascular lesions (Fazekas 1).
27	Prodromal	NA	Aβ 1-42 = 362 Aβ 1-40 = 5050 Ratio 1-42/1-40 = 0.072 t-tau = 314 (c)	Mild hypometabolism precuneus and posterior cingulate.	Mild mesial temporal atrophy

Table 8.B: Clinical biomarkers for the AD patients of the Leuven - SensorDot Alzheimer's disease (SDAD) dataset - continued

Nb	Disease stage	Amyloïd-PET	CSF Biomarkers	18F-FDG-PET	Brain MRI (or CT)
28	Prodromal	NA	NA	NA	Generalized atrophy. Prominent mesial temporal atrophy L >R. Mild white matter lesions.
29	Dementia	NA	NA	Hypometabolism temporoparietal L >R, precuneus bilateral	Generalized atrophy, parietal predominance. Mesial temporal atrophy L >R. Mild white matter lesions
30	Dementia	NA	Aβ 1-42 = 629 Aβ 1-40 = 10246 Ratio 1-42/1-40 = 0.061 t-tau = 526 (c)	Hypometabolism temporoparietal L >R and mild frontal L	CT: No significant changes
31	Dementia	Positive	NA	Hypometabolism temporoparietal (L >R), frontal (L >R), precuneus and posterior cingulate, secundary visual cortex left	Generalized atrophy. Hemosiderine depositions cortico-subcortical. Mild to moderate white matter lesions
32	Prodromal	NA	Aβ 1-42 = 370 Aβ 1-40 = 6088 Ratio 1-42/1-40 = 0.061 t-tau = 389 p-tau = 54 (c)	Mild hypometabolism posterior cingulate.	Generalized atrophy, parieto-occipital predominance. Prominent mesial temporal atrophy. Prominent white matter lesions.

Table 8.B: Clinical biomarkers for the AD patients of the Leuven - SensorDot Alzheimer's disease (SDAD) dataset - continued

Nb	Disease stage	Amyloïd-PET	CSF Biomarkers	18F-FDG-PET	Brain MRI (or CT)
33	Dementia	NA	NA	NA	Mild generalized atrophy. Mild mesial temporal atrophy.
34	Dementia	Positive	NA	Hypometabolism parieto-temporal L >R, precuneus and posterior cingulate	Mild generalized atrophy. Mild white matter lesions.
35	Prodromal	NA	Aβ 1-42 = 439 Aβ1-40 = 9506 Ratio 1-42/1-40 = 0.046 t-tau = 1173 (c)	Mild hypometabolism parietotemporal, bilateral precuneus and posterior cingulate.	Mild generalized atrophy. Mild parietal atrophy. Mild mesial temporal atrophy.

[a] Values supporting a diagnosis of Alzheimer's disease are t-tau level > 367 pg/mL, p-tau > 80 pg/mL, Aβ 1-42 < 500 pg/mL.

[b] Values supporting a diagnosis of Alzheimer's disease are p-tau > 57pg/mL, t-tau > 501 pg/mL, ratio 1-42/1-40 < 0.12, Aβ 1-42 < 775 pg/mL.

[c] Values supporting a diagnosis of Alzheimer's disease are ratio 1-42/1-40 < 0.096, t-tau > 545 pg/mL.

Amyloid PET was performed with ^{11}C-Pittsburgh compound B, except for AD043 with ^{18}F-NAV4694

Abbreviations: Aβ, Amyloid-beta; CSF, Cerebrospinal fluid; CT, computed tomography; L, left; MRI, magnetic resonance imaging; NA, not available; p-tau, tau phosphorylated at threonine 181; R, right; t-tau, total tau;

Table 8.C: Sleep staging performance for PSG and wearable data, compared between pre-trained and personalized sleep staging models. The personalized model's performance is shown on the different patient subgroups as well.

Sleep staging on PSG				
		Accuracy	κ	wF1
Pre-trained	All subjects	0.769 ± 0.011	0.665 ± 0.015	0.743 ± 0.012
Personalized	All subjects	0.808± 0.009	0.724 ± 0.014	0.801 ± 0.010
	CIE - no OSA	0.857 ± 0.010	0.799 ± 0.015	0.854 ± 0.011
	CIE - OSA	0.789 ± 0.017	0.699 ± 0.025	0.775 ± 0.019
	AD - no OSA	0.732 ± 0.024	0.607 ± 0.037	0.727 ± 0.025
	AD - OSA	0.826 ± 0.011	0.747 ± 0.018	0.823 ± 0.011
Sleep staging on wearable				
		Accuracy	κ	wF1
Pre-trained	All subjects	0.655 ± 0.014	0.498 ± 0.019	0.616 ± 0.015
Personalized	All subjects	0.703 ± 0.013	0.577 ± 0.019	0.690 ± 0.014
	CIE - no OSA	0.755 ± 0.020	0.659 ± 0.027	0.745 ± 0.020
	CIE - OSA	0.692 ± 0.025	0.568 ± 0.033	0.673 ± 0.025
	AD - no OSA	0.631 ± 0.032	0.463 ± 0.046	0.625 ± 0.033
	AD - OSA	0.694 ± 0.029	0.555 ± 0.044	0.681 ± 0.030

Performances are reported as mean ± standard error.

Extended sleep staging results

The sleep staging performance of the pre-trained model and fine-tuned, personalized model is shown in Table 8.C. The personalized model's performance is shown on the different patient subgroups as well. All performance metrics are with respect to manually scored PSG as the ground truth. For both PSG and wearable, the personalized model achieved a 4% - 5% higher accuracy compared to the pre-trained model, because it was better adapted to the individual subjects. The personalized model achieved best performance for the CIE subgroup without OSA, followed by the CIE subgroup with OSA and AD patients with OSA, which had similar sleep staging performances. The lowest sleep staging performance was observed in the group of patients with AD without OSA. Compared to the PSG-based model, the accuracy of the wearable-based model was around 10% lower.

Figure 8.A shows a few of the main clinically important sleep staging parameters in the four subgroups, derived from the manual PSG scoring and the personalized sleep staging model for the PSG and wearable. The estimations of the sleep

parameters by the PSG-based AI model were highly similar to the ground truth parameters, with three out of four significant differences between the patient subgroups reproduced. The parameters estimated by the wearable-based model differed slightly more from the ground truth, with only two overlapping significant differences. The AI's parameters based on PSG and wearable also exhibited additional significant differences.

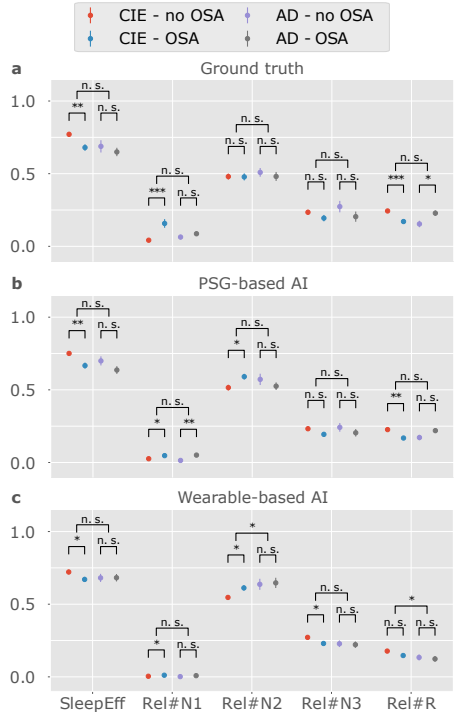


Figure 8.A: Comparison of the sleep parameters of different patient subgroups based on ground truth, PSG-based AI and wearable-based AI. The four different-colored dots each represent a patient subgroup. Sleep parameters based on (a) manually annotated PSG (ground truth), (b) AI-scored PSG, (c) AI-scored wearable data. Significant differences according to an independent two-sample t-test are indicated with * ($p < 0.05$), ** ($p < 0.005$), etc. The statistical test is performed between CIE patients with and without OSA, between AD patients with and without OSA, and between all CIE participants and all patients with AD. Abbreviations: SleepEff, sleep efficiency; Rel#X, relative time in sleep stage X (divided by total sleep time).

Extended classification and feature interpretation results

Table 8.D shows the performance in detecting patients with AD and the subgroup in the prodromal AD stage based on hypnogram and physiological features. For the hypnogram features, we compared the ground truth hypnogram features against the hypnogram features obtained from the machine learning-scored PSG, and the machine learning-scored wearable data. The physiological features were based on both the raw data and the sleep stage labels. Therefore, both the type of data and the way that it was scored is specified. The physiological features derived from the PSG data scored with ground truth sleep stages are compared to the features from the PSG data scored by the machine learning algorithm, and the features from the wearable data scored by the machine learning algorithm. For each feature set, three results are shown.

Table 8.D: Performances in detecting patients with AD and the subgroup in the prodromal AD stage based on hypnogram features and physiological features.

		Accuracy	ϰ	Precision	Recall
Hypnogram features					
Ground truth	AD	0.636 ± 0.024	0.167 ± 0.060	0.465 ± 0.041	0.406 ± 0.065
	AD dementia	0.684 ± 0.026	0.217 ± 0.066	0.400 ± 0.043	0.474 ± 0.074
	Prodromal	0.785 ± 0.023	-0.022 ± 0.060	0.131 ± 0.069	0.067 ± 0.033
PSG-based AI	AD	0.644 ± 0.014	0.165 ± 0.038	0.475 ± 0.026	0.369 ± 0.043
	AD dementia	0.711 ± 0.016	0.265 ± 0.045	0.440 ± 0.027	0.487 ± 0.058
	Prodromal	0.781 ± 0.020	-0.101 ± 0.024	0.000 ± 0.000	0.000 ± 0.000
Wearable-based AI	AD	0.664 ± 0.034	0.216 ± 0.077	0.514 ± 0.062	0.409 ± 0.054
	AD dementia	0.711 ± 0.035	0.254 ± 0.083	0.441 ± 0.063	0.461 ± 0.065
	Prodromal	0.830 ± 0.028	0.162 ± 0.056	0.439 ± 0.167	0.175 ± 0.025
Physiological features					
Ground truth (PSG)	AD	0.857 ± 0.023	0.676 ± 0.051	0.818 ± 0.040	0.751 ± 0.036
	AD dementia	0.897 ± 0.024	0.738 ± 0.060	0.770 ± 0.049	0.852 ± 0.048
	Prodromal	0.847 ± 0.013	0.339 ± 0.052	0.502 ± 0.059	0.375 ± 0.067
PSG-based AI	AD	0.822 ± 0.021	0.588 ± 0.048	0.788 ± 0.037	0.657 ± 0.038
	AD dementia	0.881 ± 0.019	0.695 ± 0.046	0.750 ± 0.040	0.804 ± 0.035
	Prodromal	0.846 ± 0.024	0.328 ± 0.089	0.497 ± 0.104	0.358 ± 0.065
Wearable-based AI	AD	0.797 ± 0.022	0.544 ± 0.049	0.717 ± 0.035	0.677 ± 0.048
	AD dementia	0.832 ± 0.023	0.581 ± 0.054	0.650 ± 0.042	0.752 ± 0.052
	Prodromal	0.832 ± 0.018	0.234 ± 0.060	0.427 ± 0.083	0.267 ± 0.050

AD: the performance of a machine learning model trained to discriminate AD patients, including those in the prodromal disease stage.
AD dementia: results of the previous model, but only including patients in the dementia stage for performance testing.
Prodromal: the performance when training a model to classify the prodromal AD patients from cognitively intact elderly.
The results are reported as mean ± standard deviation.

To explore OSA's impact as a confounder on our AD detection model, Figure 8.B depicts the AD detection performance for subjects with and without OSA separately. The ROC curves show that the discrimination power for AD patients with and without OSA was not very different in most cases. Similarly, Figure 8.C illustrates the impact of age as a confounding factor on the AD detection model, comparing its performance for younger and older AD patients. Again, the discrimination power for younger and older patients is similar, indicating that the AD model is truly detecting AD rather than age.

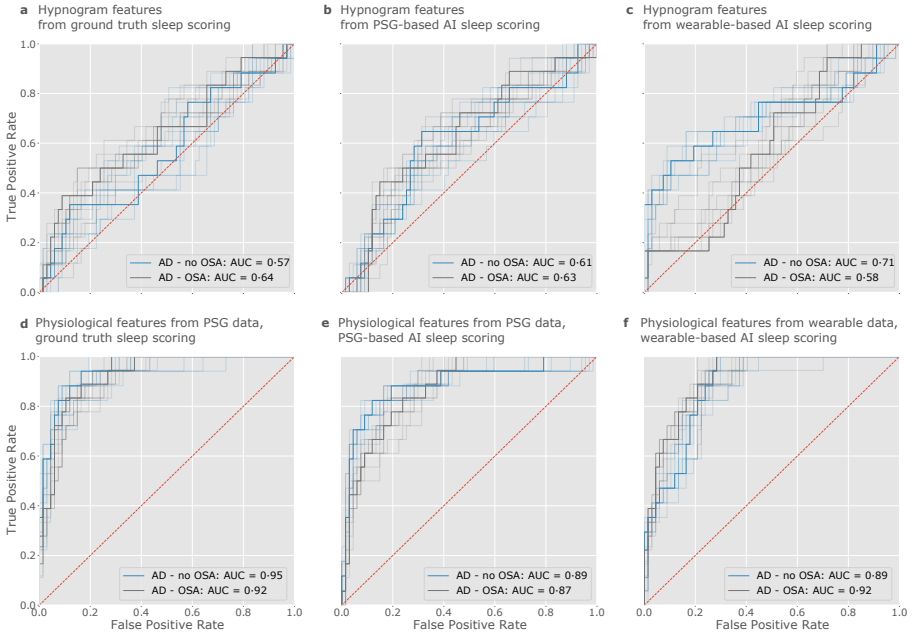


Figure 8.B: ROC performance curves for detecting AD with and without OSA. The transparent lines represent ROC curves of all ten trained classifiers, while the non-transparent lines show ROC curves using the mean predictions from ten repeats, with AUC reported in the legend. Curves labeled with 'AD - no OSA' assess AD discrimination in participants without OSA (AHI $\geq$ 15), while 'AD - OSA' curves assess AD detection in participants with OSA. Both results used the same trained models, with only the test data differing. The upper row shows performances using hypnogram features, calculated based on (a) manual PSG labels (ground truth), (b) AI-labeled PSG, (c) AI-labeled wearable data. The lower row shows performances using physiological features, calculated using both the raw data and sleep stage labels: (d) raw PSG data with manual labels, (e) raw PSG data with AI-based labels, and (f), wearable data with AI-based labels.

A last confounding factor to investigate is the impact of the dataset. Two

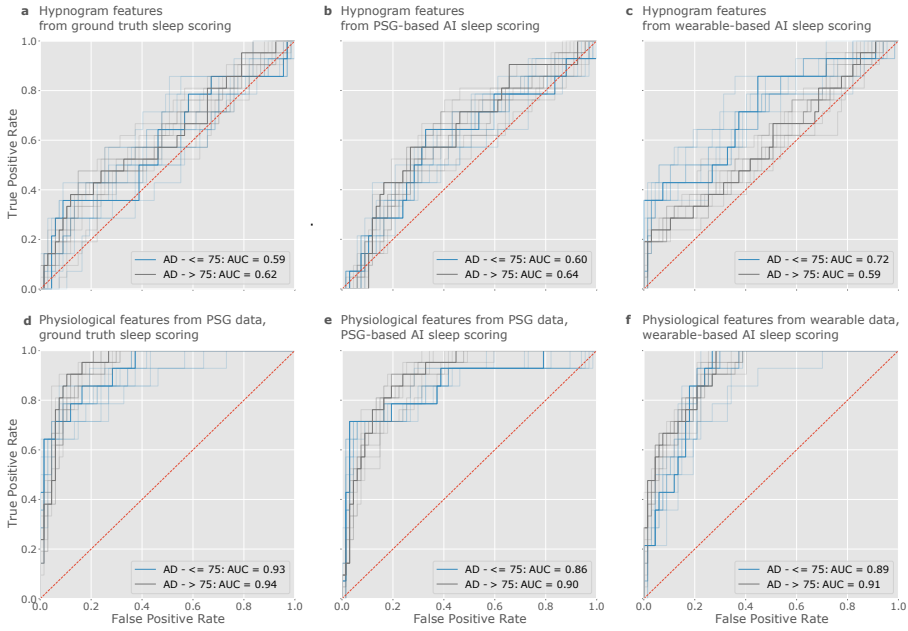


Figure 8.C: ROC performance curves for detecting AD in younger and older subjects. The transparent lines represent the ROC curves of all ten classifiers trained for each prediction task, while the non-transparent lines represent the ROC curves obtained using the mean predictions with ten repeats. Their AUC is reported in the legend. The curves labeled with 'AD - $\leq$ 75' show the performance when testing how well the AD detection model could detect AD patients up to 75 years old (n = 14, mean age = 68.4), and the curves labeled with 'AD - >75' show the performance when testing how well AD patients older than 75 were discriminated (n = 21, mean age = 78.8). In both cases, the negative group consists of all 67 non-AD subjects. Both results were computed with the same trained models, with only the test data differing. The upper row shows the performances using hypnogram features, calculated based on (a) manual PSG labels (ground truth), (b) AI-labeled PSG, (c) AI-labeled wearable data. The lower row shows performances using physiological features, calculated using both the raw data and sleep stage labels: (d) raw PSG data with manual labels, (e) raw PSG data with AI-based labels, and (f), wearable data with AI-based labels. The ROC curves show that the discrimination power for AD patients with older age usually distinguished slightly better, but that the performance for both age groups was not very different in most cases.

different datasets were merged for this study, and so, it is relevant to assess whether we are able to discriminate AD patients from non-AD patients from both datasets. Figure 8.D shows the model's power in discriminating AD patients from CIE from each of the two datasets. The ROC curves reveal that

the discrimination power is comparable for both datasets when using hypnogram features, but significantly higher for the SDS dataset when physiological features are used. This indicates that the signals might contain subtle information distinguishing the datasets, even though both were collected using nearly identical measurement setups. Despite this, the model's ability to differentiate between AD and CIE patients in the ADSD dataset remains strong.

Additionally, we investigated how much discriminatory power lies in the clinical data (the questionnaires) alone. The PSQI and ESS questionnaires investigate sleep quality of the patient, subjectively, and the MMSE questionnaire is used to test cognitive skills. We used these three scores and the age of the participants to predict AD and prodromal AD with an XGBoost model and compared the resulting prediction performance to that of the MLP model based on PSG data. Since the questionnaires were only available for the SDAD dataset, this experiment is performed purely on that dataset. Figure 8.E shows the results. XGBoost based on clinical data demonstrates significant discriminatory power compared to the MLP based on PSG data. It performs marginally worse for AD but slightly better for prodromal. When the two classifiers are combined in an ensemble model, the performance surpasses that of either model individually. This experiment suggests that signals acquired during sleep can potentially also be combined with clinical questionnaires and patient characteristics to more accurately diagnose or predict AD.

Figure 8.F shows the correlation between MMSE score and sleep features based on the PSG data, indicating a strong correspondence between the manually scored and AI-scored PSG features.

Figure 8.G shows the physiological features selected in more than 85% of iterations, both based on AI scoring and ground truth scoring, and both for the wearable device and PSG. The imperfections in sleep staging are apparent when comparing Figure 8.Ga to Figure 8.Gb, and Figure 8.Gc to Figure 8.Gd, especially for the wearable features. For instance, frequency bands that were important in N2 and N1 according to ground truth scoring (Figure 8.Ga) appear important in REM when scored by AI (Figure 8.Gb). Nonetheless, this analysis also demonstrated that despite imperfect sleep staging by AI, the important features remained identifiable and valuable for AD detection. This is evidenced by the consistent importance of certain frequencies with AI scoring, albeit in somewhat different sleep stages.

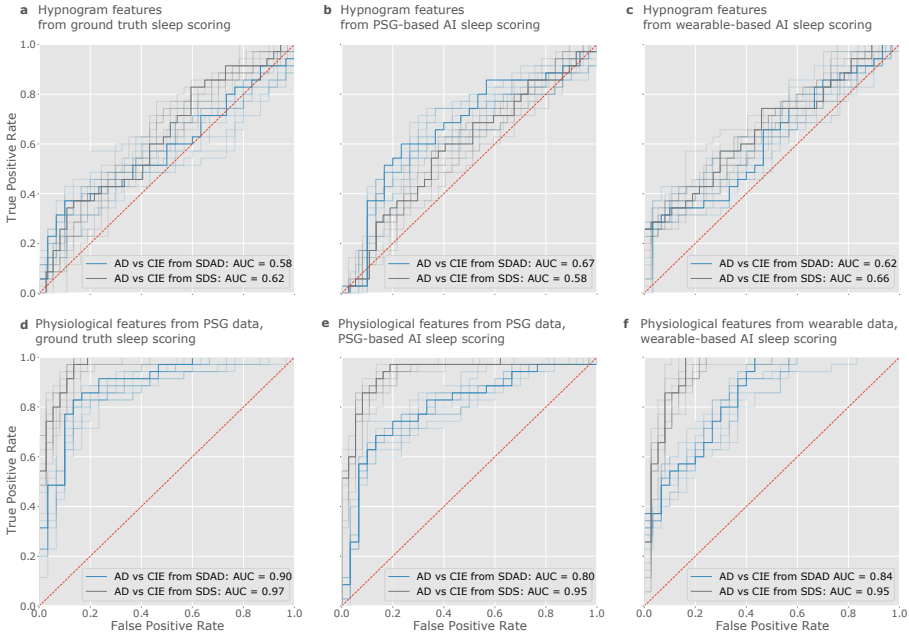


Figure 8.D: ROC performance curves for discriminating AD from CIE from each of the two datasets. The transparent lines represent the ROC curves of all ten classifiers trained for each prediction task, while the non-transparent lines represent the ROC curves obtained using the mean predictions with ten repeats. Their AUC is reported in the legend. The curves labeled with 'AD vs CIE from SDAD' show the performance when testing how well the AD detection model discriminated AD patients from CIE subjects in the SDAD study, and the curves labeled with 'AD vs CIE from SDS' shows the discrimination power compared to CIE subjects from the SDS study. In both cases, the positive group consists of all 35 AD patients from the SDAD dataset. Both results were computed with the same trained models, with only the test data differing. The upper row shows the performances using hypnogram features, calculated based on (a) manual PSG labels (ground truth), (b) AI-labeled PSG, (c) AI-labeled wearable data. The lower row shows performances using physiological features, calculated using both the raw data and sleep stage labels: (d) raw PSG data with manual labels, (e) raw PSG data with AI-based labels, and (f), wearable data with AI-based labels. The ROC curves show that the discrimination power for the two datasets is similar with the hypnogram features, but slightly better for the SDS dataset with the physiological features.

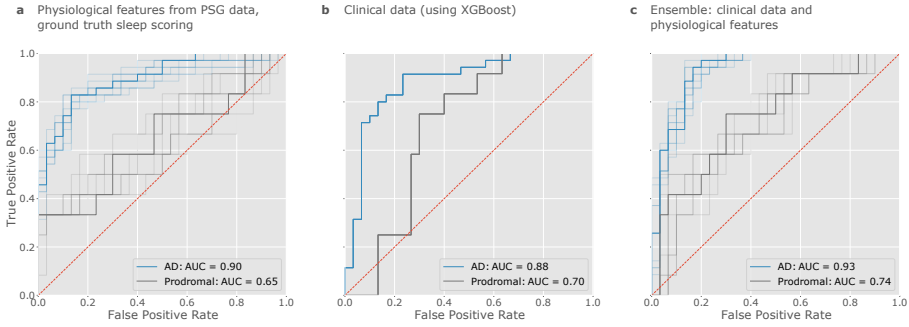


Figure 8.E: ROC performance curves for discriminating AD and prodromal AD from CIE using physiological features and using clinical data from questionnaires. The experiment was performed using only the SDAD dataset (since the questionnaires were only available for that dataset). The clinical data used for the XGBoost model is the MMSE, PSQI, ESS and age of the participant. The transparent lines represent the ROC curves of all ten classifiers trained for each prediction task, while the non-transparent lines represent the ROC curves obtained using the mean predictions with ten repeats. Their AUC is reported in the legend. As the prodromal AD patients are a subgroup of AD, the curves labeled with 'AD' show the performance in discriminating CIE patients from the whole AD cohort, consisting of patients in the prodromal and dementia disease stage. The curves labeled with 'prodromal' show the performance in discriminating the prodromal AD patients from the CIE group. Both results were computed with the same trained models, with only the test data differing. The classification performances are calculated based on (a) raw PSG data with manual labels, (b) clinical data with XGBoost as a classifier (f), both the XGBoost clinical data model and the PSG-based MLP model. The ROC curves show that XGBoost based on clinical data demonstrates significant discriminatory power. Compared to the MLP based on PSG data, it performs marginally worse for AD but slightly better for prodromal. When the two classifiers are combined in an ensemble model, the performance surpasses that of either model individually.

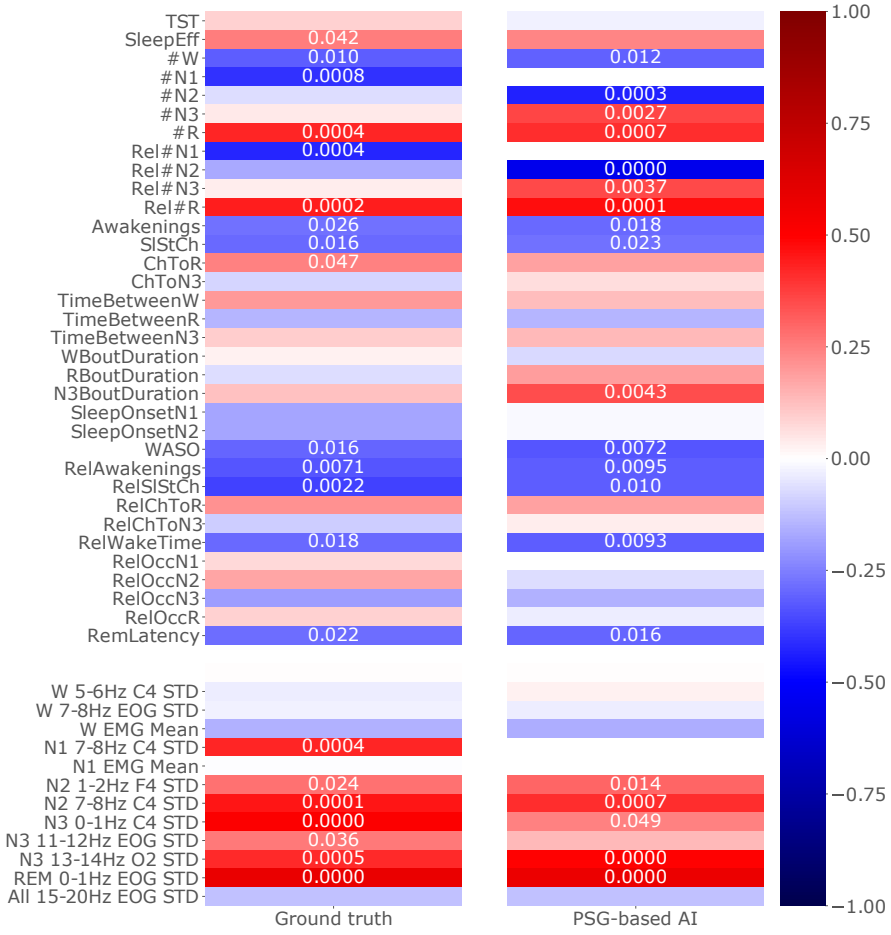


Figure 8.F: Correlation between sleep features based on the PSG data and MMSE. The hypnogram features are shown in the upper part of the figure, and a few selected physiological features are shown in the lower part. The first column shows the ground truth sleep features based on PSG data with manual scoring, and the second column shows sleep features based on PSG data scored by AI. Pearson's correlation test was performed with the subjects as samples. The MMSE score was not recorded in the patients from the SDS dataset, so the correlations were only computed for the 65 patients of the SDAD dataset (see Table 8.3). Correlations are shown through the colors, with p values below 0.05 reported as numbers to indicate the significance of the correlations. The PSG-based AI agrees on most significant correlations that are found in the ground truth, with an agreement of 73.5% for the hypnogram features and 88.8% for the physiological features. Abbreviations: All, all sleep stages; STD, standard deviation. Abbreviations for hypnogram features are explained in Table 8.A.

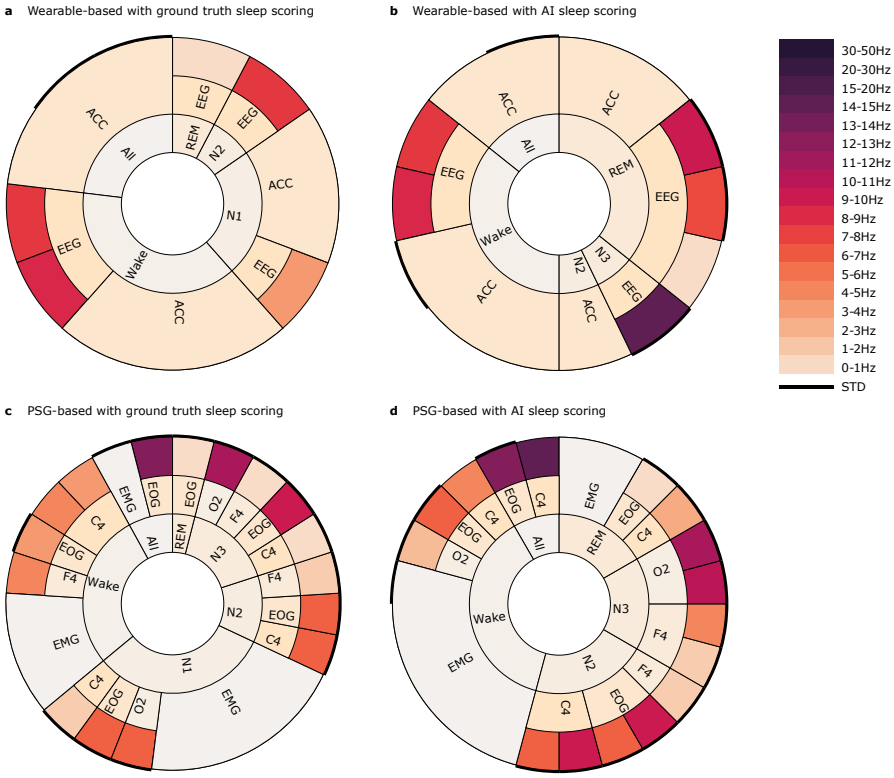


Figure 8.G: Most selected physiological features based on ground truth scoring and AI scoring. The physiological features that were selected more than 85% of the time by (a) the wearable-based AD detection model using ground truth sleep scoring, (b) the wearable-based AD detection model using AI sleep scoring, (c) the PSG-based AD detection model using ground truth sleep scoring, (d) the PSG-based AD detection model using AI sleep scoring. Features are visualized using three rings showing to which (1) sleep stage, (2) channel, and (3) frequency range each feature corresponded. The frequency range is only shown for the EEG and EOG channels due to the limited interpretability of the EMG and ACM frequencies. The discrete color bar shows the interpretation of the frequency range. The surrounding black ring shows which features were computed using STD, with the remaining calculated using the mean. The width of each segment indicates how many features of that type were selected more than 85% of times. This width should not be mistaken for overall importance of a channel or sleep stage: information may be more spread out over multiple frequency bands for some channels or sleep stages than for others. Rather, this figure shows which channels were important in which sleep stages, and which frequencies were decisive in each channel. Abbreviations: All, features computed over the whole recording.

Part V

Conclusion

9 | Conclusion

This thesis contributes to the development of automated sleep staging methods, addressing the difficulties that limit its wide adoption for home-based and in-hospital sleep studies. The three main contributions are detailed in the previous three parts. This concluding chapter summarizes the main findings of this work and discusses avenues for future research.

9.1 Main findings

9.1.1 Part II - Transfer learning approaches for wearable EEG data

The first main part of this thesis (Part II: Chapter 3-4) addresses its first subobjective. It introduces methods that improve the generalization performance of sleep staging models, aiming to achieve better sleep staging performances on wearable EEG device measurements. As labeled data are expensive to acquire with wearable devices, we need (a) methods that can learn from small amounts of labeled data, or (b) methods that can learn from unlabeled data. I devoted one chapter to each of the two approaches.

Chapter 3 presents a solution utilizing small amounts of labeled data. It presents a novel supervised transfer learning technique for sleep staging on wearable devices. The method is based on feature matching, which aligns the features of paired samples from a source and target domain. This approach effectively improves sleep staging performance on new wearable devices or sleep measurement setups with only a few labeled recordings. The method requires simultaneous measurements of the source and target modalities, a condition that is automatically met when labels are available in sleep recordings from

wearable device data. With smart use of those simultaneous recordings, the feature matching approach outperforms fine-tuning, the standard supervised transfer learning approach, in multiple sleep staging tasks.

Chapter 4 focuses on utilizing unlabeled data. It presents an adversarial domain adaptation approach, and investigates several crucial technical aspects to apply this approach to sleep staging. The method proves useful in sleep staging on wearable device recordings, and the addition of pseudo-labels or real target domain labels can further improve the performance. Additionally, minimizing domain mismatch by choosing a source domain similar to the target domain also improves performances. In sleep datasets, aspects that influence this similarity are the scorer, acquired channels, equipment, and studied population. Lastly, unsupervised adversarial domain adaptation even shows small performance improvements in recording-wise adaptation (or 'personalization'), an extreme version of domain adaptation where the target domain is a single recording.

The methods and learnings from Part II not only demonstrate their value in the use cases outlined in these chapters, but also show their utility in clinical applications discussed in Part IV of this thesis. For sleep scoring of labeled PSG datasets of a reasonable size, I opted for simple fine-tuning in the Parkinson's disease dataset (Section 7.2) and the pediatric epilepsy dataset (Section 7.3). I deployed the power of unlabeled sleep recordings with adversarial domain adaptation in the adult epilepsy SensorDot dataset (Section 7.4) and the AD DREEM2 dataset (Section 7.5). For both these datasets, I applied the insight from Chapter 4 concerning minimizing the domain mismatch. Accordingly, the respective selected source domains were the SDAD SensorDot dataset (CIE portion) and the Dreem - Headband dataset, utilizing the same devices and channels as the target domains to minimize the domain mismatch. Lastly, in Chapter 8, a combination of fine-tuning and unsupervised recording-wise domain adaptation achieved the best sleep staging performances on the SDAD dataset.

In conclusion, Part II discusses methods to address challenges related to performance and generalizability of automated sleep staging tools. These methods were proven useful in various clinical populations, and for both hospital PSG and wearable devices.

9.1.2 Part III - Uncertainty estimation to increase trust

After addressing the challenges related to model performance in Part II, Part III (Chapter 5-6) delves into another critical aspect impacting the adoption of automated decision support systems in clinical practice: trust and transparency.

This part investigates enhancing the transparency of deep learning models for automated sleep staging by incorporating uncertainty estimation.

In Chapter 5, an uncertainty-guided pipeline for automated sleep staging (U-PASS) is developed to optimize reliability and usability of automated sleep staging aids in clinical practice. This model-agnostic framework improves model reliability with three mechanisms: first, by selecting or excluding training data based on uncertainty measures; second, by fine-tuning a model using a clinician-in-the-loop active learning approach with uncertainty-based sampling; and third, by deferring uncertain predictions to an expert. While the first step purely improves model performance, the latter two steps merge model performance and subjective user experience. By including the clinician in the decision-making process, they are more likely to trust and use this decision support system. Although U-PASS is presented as a pipeline, individual steps from this framework can be integrated and combined with any machine learning model and problem.

An essential contribution of this chapter is the differentiation between data uncertainty and model uncertainty. This distinction allows us to discern between intrinsic, unavoidable uncertainties, such as those arising from imperfect scoring rules, and uncertainties that can be mitigated through additional training. Model uncertainty, reflecting a lack of knowledge from the model, can emerge from domain shifts, which can be addressed through transfer learning (Part II). The uncertainty-based active learning approach proposed to fine-tune the model and mitigate this uncertainty can thus be used as a transfer learning approach. We applied the semi-supervised adversarial domain adaptation approach from Chapter 4, with labels obtained by querying a clinical expert. With Chapter 3 and Chapter 4 proposing methods to tackle domain shift, this chapter provides tools to diagnose it.

Chapter 6 applies the U-PASS pipeline to sleep staging with wearable devices, exploring how uncertainty estimation could be useful for at-home sleep monitoring. Through experiments involving manual review by a clinician, we confirm the difficulties experienced by human scorers in interpreting the signals from a wearable behind-the-ear device and in correcting uncertain predictions from the model. As a result, U-PASS needs to be adjusted for this application: while training data curation and flagging of uncertain data at deployment time are feasible, human feedback in active learning or correction at deployment time are not. Since it is relatively cheap and easy to perform a new measurement with a wearable, we propose discarding recordings with high overall uncertainty. Regarding epoch-level uncertainty, we demonstrate that most sleep parameters, aggregating information over an entire night, remain relatively well predicted, and reliably reflect known correlations with a clinical parameter. Consequently, we conclude that large recording-level uncertainty can be remedied by repeating

measurements, and small epoch-level uncertainties are not so problematic for the purpose of large-scale or long-term sleep measurements across the population.

9.1.3 Part IV - Validation in clinical populations

Circling back to the challenge identified in Section 1.4 regarding the clinical validity of sleep staging algorithms, Part IV (Chapter 7-8) validates these algorithms in different clinical populations.

Chapter 7 summarizes the results of different collaborations in which I applied automated sleep analysis methods, especially those developed in Part II, to different clinical populations and applications. This chapter covers sleep staging tasks for PSG and wearable recordings in patients with epilepsy, AD and PD. The findings show that it is possible to characterize sleep in these populations, that known clinical correlations and findings can be reproduced with the derived parameters, and that sleep staging even has a use in therapeutic applications. The transfer learning methods manage to effectively boost sleep staging accuracies, with different approaches proving useful in different scenarios. For PSG datasets, fine-tuning a neural network pre-trained on another PSG dataset is generally effective. For wearable datasets, if labeled training data are available, the feature matching approach is preferable as it leverages information from paired PSG and wearable measurements. When dealing with large amounts of unlabeled data, semi-supervised or unsupervised domain adaptation approaches should be used, depending on the availability of labeled data.

Chapter 8 goes beyond sleep staging and provides an in-depth study of sleep as a diagnostic marker for AD. Changes in sleep are related with a wide range of psychiatric and neurological conditions, making non-invasive sleep monitoring a promising tool for (early) detection or longitudinal follow-up of these disorders. This chapter investigates the use of sleep measurements to screen for AD, finding that automatically analyzed EEG data from wearables indeed presents a promising approach for large-scale, early AD detection. Perfect five-stage sleep staging with expert accuracy is not necessary for this purpose, and the frequency content of the signals averaged by sleep stage (microstructural sleep alterations) holds more diagnostic information than typical sleep stage features (macrostructural).

Addressing the question ‘when is the performance of automated sleep staging algorithms good enough?’ (see Section 1.4), the answer depends on the application. Sleep staging itself is seldom the desired end goal of a sleep study, so the desired accuracy depends on the intended use of the resulting sleep stages. If the sleep stages are used to derive sleep parameters, performance

should be evaluated by comparing to reference parameters. If the goal is to detect or diagnose a disorder, the detection performance should be compared between AI scoring and manual PSG scoring. In therapeutic applications like the DREEM2 headband for deep sleep stimulation, the AI's performance should be sufficient to achieve the desired therapeutic outcome. As shown in Chapter 7 and Chapter 8, when enhanced with transfer learning approaches, state-of-the-art sleep staging algorithms achieve adequate accuracy to be used in clinical practice for both PSG-based and wearable-based sleep staging. Additionally to demonstrating sufficient accuracy and clinical validity on a 'population' level, the uncertainty quantification methods outlined in Part III provide extra safety railguards to ensure the expected performance is always met and that problematic recordings or segments are accurately identified.

9.2 Future research directions

9.2.1 Open technical challenges in automated sleep staging

Domain adaptation: how much and when?

In relation to the challenge of improving sleep staging performance for wearable sleep monitoring, further research is needed to thoroughly investigate when to use unsupervised domain adaptation and whether it is best performed at a population level, subject level or recording level. Computationally, population-level adaptation is the most advantageous, followed by subject-level adaptation and then recording-level adaptation. In a recent collaboration, we found that unsupervised adversarial domain adaptation did not consistently improve performance with subject-wise adaptation when using a few good-quality, pre-cleaned wearable device recordings of healthy subjects [1], whereas supervised fine-tuning achieved consistent improvements. Similarly, in the adult epilepsy population (Section 7.4), unsupervised domain adaptation did not improve results for standard EEG recordings but did improve performance for wearable data, both for recording-wise and subject-wise adaptation. We conclude that unsupervised domain adaptation is particularly useful when baseline performances are low. Further research is needed to establish whether subject-wise adaptation is feasible with unsupervised domain adaptation, although I strongly suspect that this is the case with a large enough number of (unlabeled) recordings of a subject. Determining the exact number of recordings required and

[1] K. P. Lorenzen, **E. R. M. Heremans**, M. De Vos, and K. B. Mikkelsen, "Personalization of automatic sleep scoring: How best to adapt models to personal domains in wearable EEG," Accepted in *IEEE Journal of Biomedical and Health Informatics*, 2024.

how this varies with limited labeled data remains an open question. Additionally, some populations (especially healthy ones) may not require personalization and might only need population-level adaptation, whereas populations with more variations in sleep patterns and disturbances may benefit from more personalized sleep staging approaches.

Clinician-in-the-loop for wearables

In our work on uncertainty quantification for wearables (Chapter 6), the inability of a clinician to improve the machine learning system's sleep stage predictions on wearable EEG sparked debate among the co-authors. Although this study cannot be generalized to all clinicians and wearables, it highlights an important issue with wearables, for which I envisage two possible solutions. One possible way forward is to perform fully automatically AI-scored sleep studies with wearables, without sleep stage verification by clinicians. This would require strong performance guarantees for wearables and AI solutions, and clinicians would need to trust the outcomes. Solutions like U-PASS can help build this trust. The other possible way forward is to further improve the signal quality and transform the EEG signals from wearables to better resemble standard EEG signals. Future research should investigate if transformer or RNN architectures can 'translate' one EEG derivation to another one, allowing clinicians to apply their sleep staging expertise based on known EEG derivations to interpret signals and correct AI-based scoring as needed. The fully automated scoring approach without clinician intervention may be suitable for large-scale screening purposes and long-term monitoring, whereas a clinician-in-the-loop approach with PSG-like signal quality may be needed for detailed diagnostic sleep studies.

Transparency beyond uncertainty estimation

Uncertainty estimation is not the only way to increase transparency of deep learning models. The field of explainable artificial intelligence (XAI) has recently gained significant attention [182, 183], aiming to provide more insight into how models make decisions. This could be hugely beneficial to ensure effective collaboration between experts and AI in the clinic. Post-hoc explainability methods are especially popular as they provide explanations without compromising the models' predictive power. A few works have made attempts to integrate XAI into sleep staging methods [50, 52, 122, 184, 185], but there is still considerable potential for further investigation. The true challenge lies in evaluating these explainability methods' performance and understanding which types of explanations clinicians prefer.

Another avenue for future research is to explore the intersection between uncertainty estimation and explainability. Simply quantifying the level of uncertainty may not always be sufficient, as a user may wonder why this uncertainty arises. Is the sample noisy, does it contain features of multiple classes, or is it out-of-distribution? Post-hoc explainability methods could be used to explain the uncertainty associated with individual predictions. This could help doctors to understand why an AI abstains from making a prediction or indicates high uncertainty. We conducted proof-of-concept experiments to see if explainability methods like feature occlusion [186] or SHAP [187] could be harnessed to explain sources of uncertainty (see Appendix A). The results of this preliminary work suggest that this approach could indeed be beneficial for automated sleep staging aids.

9.2.2 Future perspectives: AI propelling sleep research forward

Sleep analysis beyond sleep staging

The main focus of this thesis was sleep staging according to the AASM guidelines [7]. There are a number of problems with these guidelines, most notably their inherent ambiguity, which is evidenced by the considerable inter-rater disagreement [121, 142]. Much of this ambiguity arises from scoring sleep in 30-second epochs, a practice rooted in the days when sleep records were printed on paper and not at all based on the biology of sleep. Shifting from 30-second segmentation to a more fine-grained approach may reduce disagreement, as multiple stages present in a single 30-second epoch could be scored more accurately. Going beyond more fine-grained sleep staging, recent works have even advocated for abolishing sleep stages altogether, arguing that modeling sleep as a continuous process would provide a more biologically accurate representation [143–145]. Machine learning methods, like the uncertainty metrics presented in this thesis, could help define and describe sleep using new, more biologically plausible models.

Large-scale sleep monitoring

Lastly, sleep research would greatly benefit from more large-scale sleep studies. This thesis presents proof-of-concepts demonstrating the clinical value of sleep measurements on small, tailored patient samples. However, to truly capture and understand the intricate dynamics between sleep and various disorders on a population level, we need large-scale longitudinal studies that monitor sleep, diagnose disorders, and track other clinical parameters over time across large

populations. Such studies would enable us to evaluate the effectiveness of using sleep measurements with AI for disease screening and tracking the progression of disorders through sleep-related markers.

Wearables and AI have the potential to be used for both the diagnosis and early detection of disorders, as well as for prognosis, aiding in tracking the disease progression of disorders related to sleep. The first population I would target for large-scale screenings would be older adults, where sleep disturbances are especially common and related to neurodegeneration. One or a few nights of recordings per participant would likely contain sufficient information for large-scale disease screening. For long-term tracking of disease progression, I would target populations with diagnosed neurological or neurodegenerative illnesses, measuring sleep over several months. This thesis has demonstrated that wearables can assist in detecting AD, but much larger sample populations are needed to investigate the ability to detect and differentiate between other neurodegenerative diseases.

This thesis provides strong evidence of the many possibilities of wearables and AI for sleep monitoring, but these advancements now need to be implemented on a larger scale. Only then can we truly understand sleep's relation with many disorders and learn how to use sleep as a potential marker or even a therapeutic target for these related disorders. In this vision of the future of sleep medicine, we can truly reap the benefits of these novel technologies.

9.3 Final thoughts

By tackling challenges related to integrating AI into daily sleep medicine and preparing for future wearable at-home sleep studies, this thesis has contributed several crucial elements toward making automated sleep analysis a part of clinical practice. Section 9.2.1 demonstrated that some technical challenges still need to be addressed in order to make automated sleep staging methods clinic- and future-proof. Conversely, Section 9.2.2 discussed AI's huge potential to transform sleep research and sleep medicine fundamentally.

Personally, I believe the role of AI in future sleep medicine is two-fold. Large-scale screening efforts and long-term follow-up with wearables will warrant fully-automated sleep analysis. On the other hand, clinical diagnosis of sleep disorders will require hospital-based PSG or PSG-like signals from high-quality wearables, analyzed with closed-loop interaction between clinicians and AI systems. With these two approaches, AI can help detect sleep problems in the general population without compromising the precision of sleep disorder diagnosis.

Appendices

A | Explaining uncertainty in AI

This appendix is preliminary research, building on **E. R. M. Heremans** and M. De Vos, "Explaining uncertainty in AI for clinical decision support systems," Accepted in Proceedings of the *Uncertainty meets Explainability Workshop at ECML-PKDD 2023*, Torino, Italy, pp. 1-9, 2023.

ABSTRACT | The use of AI methods in critical areas like autonomous driving and healthcare has sparked concerns about their trustworthiness and reliability. In deep learning, uncertainty quantification has been extensively studied as a potential approach to enhance trust and reliability, by measuring an AI system's confidence in its predictions. While various techniques exist for estimating uncertainty in machine learning predictions, few delve into the underlying reasons. These reasons can partly be unraveled by distinguishing between two primary types of uncertainty: epistemic uncertainty, stemming from model limitations, and aleatoric uncertainty, inherent to the data itself. This distinction is crucial for practitioners relying on AI-driven decision support. Furthermore, it would be valuable to understand which inputs contribute to the uncertainty. XAI encompasses methods aimed at shedding light on AI predictions, revealing which inputs influenced the model's decisions. This study focuses on the relatively uncharted territory where XAI and uncertainty quantification intersect, by utilizing established XAI techniques to explain the uncertainty associated with machine learning models. Specifically, we investigate the use of post-hoc feature attribution methods to compute explanations for epistemic and aleatoric uncertainty within deep neural networks.

A.1 Introduction

To our best knowledge, explaining the uncertainty of AI decisions is a largely unexplored area. Simply quantifying the level of uncertainty does not always give enough information, as a user may wonder why this uncertainty arises: is the sample noisy, does it contain features of multiple classes, or is it out-of-distribution? We envisage exploiting post-hoc explainability methods to explain the uncertainty associated with those individual predictions. In healthcare settings, this could help doctors to understand why an AI abstains from making a prediction or indicates a high level of uncertainty. We have performed proof-of-concept experiments for this idea on a synthetic dataset and on the MNIST dataset. In future work, we envision translating these methods to sleep staging.

A.2 Uncertainty explanations

Our uncertainty explanation methodology essentially consists of two steps

1. Quantify the uncertainty (both epistemic and aleatoric)
2. Explain both uncertainty measures

In order to quantify the uncertainty, we rely on Bayesian dropout uncertainty as defined by Gal and Gharimani [155]. This work introduces dropout as a way to approximate Bayesian inference to measure uncertainty in deep neural networks. We then rely on the decomposition of uncertainty into aleatoric and epistemic terms, an idea introduced by Depeweg et al. [188] and further developed by Shaker and Hüllermeier [141]. We apply this decomposition method to dropout uncertainty, resulting in the following uncertainty estimations:

$$\begin{aligned} v_{al} &= \mathbb{E}_{\Theta}\big[H[p(y|x,\Theta)]\big] \\ &\approx \frac{1}{M}\sum_{i=1}^{M} H[p(y|x,\theta_i)] \end{aligned} \tag{A.1}$$

,

$$\begin{aligned} v_{ep} &= H\big[\mathbb{E}_{\Theta}[p(y|x,\Theta)]\big] - \mathbb{E}_{\Theta}\big[H[p(y|x,\Theta)]\big] \\ &\approx H\Big[\frac{1}{M}\sum_{i=1}^{M} p(y|x,\theta_i)\Big] - \frac{1}{M}\sum_{i=1}^{M} H[p(y|x,\theta_i)] \end{aligned} \tag{A.2}$$

,

where $H = -\sum_{y=1}^{Y} p(y|x, \theta_i) \log(p(y|x, \theta_i))$ is the Shannon entropy of the predictive posterior distribution. θ_i is an instantiation of the model parameters with dropout. We can thus make approximations of uncertainty by evaluating the model multiple times with dropout. In our experiments, we used $M = 20$ evaluations.

In order to explain the uncertainty, we used SHAP values as our post-hoc explanation method. Specifically, we deployed the Deep SHAP algorithm [187], which approximates SHAP values using a selection of background samples.

A.3 Experimental methodology

In this section, we outline our experimental approach, which involves experiments on the synthetic two-moons dataset and the MNIST image dataset.

A.3.1 Synthetic dataset: two moons

Dataset description

The first experiment includes the two moons dataset in Scikit-Learn [181]. A dataset was generated with 50000 points with noise factor 0.3 and random state 0. 1000 points were used as a test dataset, and from the remaining points, the train dataset was constrained to be a circle with radius 1 and centre (0.5, 0.25), which amounts to approximately one third of the dataset. The dataset consists of two classes, and the binary classification task consists of determining which point belongs to which class. The two moons dataset with the delimiting circle is shown in Figure A.1.

Model architecture

The model architecture was a two-layer neural network with a hidden layer of 512 neurons, trained for 15 training epochs. Dropout was used before both layers, and the hidden layer was followed by a ReLU activation function. The output layer used a softmax activation. The ADAM optimizer [189] was used to optimize the cross-entropy loss function with a learning rate of 1e-3.

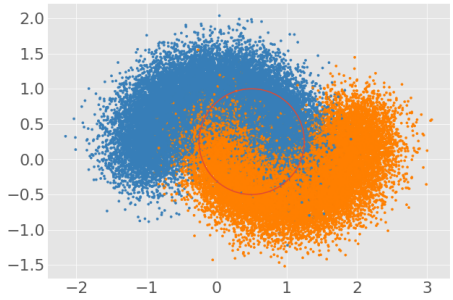


Figure A.1: The synthetic two moons dataset, which forms a binary classification problem. The color of a point indicates its class. The training set only contains points within the red circle, while the test set contains points within and outside the circle.

SHAP value calculation

We used 100 random samples from the train set as background samples for Deep SHAP [187]. We calculated the SHAP values for all 1000 test set samples.

Epistemic uncertainty explanations

We used SHAP values to explain the contribution of each input feature to the estimated epistemic uncertainty (A.2). First, we verified the epistemic uncertainty explanations with SHAP dependence plots showing how both inputs, x_1 and x_2 contributed to the uncertainty. Then, we verified the SHAP values' ability to discern which feature was more out-of-distribution, by comparing which SHAP value was larger to which feature lay further from the training distribution's centre.

Aleatoric uncertainty explanations

Like with the epistemic uncertainty, we used SHAP values to explain the contribution of both input features to the estimated aleatoric uncertainty (A.1). The SHAP values were approximated with the same approach. First, SHAP dependence plots were used to show how both inputs, x_1 and x_2 contributed to the uncertainty. Then, we verified that our SHAP values also reflect the negative relationship between aleatoric uncertainty and model confidence. Features that contribute to a high aleatoric uncertainty should not contribute much to the prediction, and, conversely, features that contribute to the prediction should not contribute to aleatoric uncertainty.

Table A.1: CNN architecture used for MNIST dataset

Type	Nb in	Nb out	Kernel	Probability
Conv	1	10	5x5	
Max pool			2x2	
ReLU				
Conv	10	20	5x5	
Dropout				0.5
Max pool			2x2	
ReLU				
Fully connected	320	50		
ReLU				
Dropout				0.5
Fully connected	50	10		
Softmax				

Abbreviations: Conv: convolutional layer; Nb in: number of inputs (channels in the case of convolutional layers, features in case of fully connected layers); Nb out: number of outputs; Prob: probability

A.3.2 MNIST dataset

Dataset description

The second experiment was performed on the MNIST dataset [190], containing 28×28 size images of handwritten digits from 0 to 9. The training set consists of 60000 images and the test set consists of 10000 images. The multi-class classification task consists of determining which digit is depicted in each image. The pixel values in the range [0,255] are converted to values in the range [0.0,1.0].

Model architecture

A CNN was selected for this classification task. The architecture consists of two convolutional layers, followed by two fully connected layers. The architecture is detailed Table A.1, and trained for 10 epochs with a batch size of 128. The SGD optimizer is used with a learning rate of 0.01 and a momentum of 0.5.

SHAP value calculation

Like for the two moons dataset, we used 100 random samples from the train set as background samples for Deep SHAP [187].

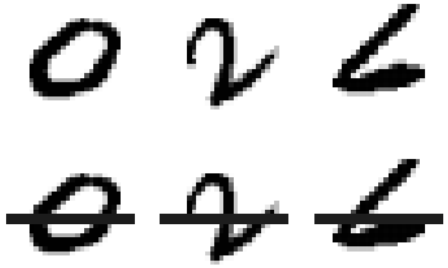


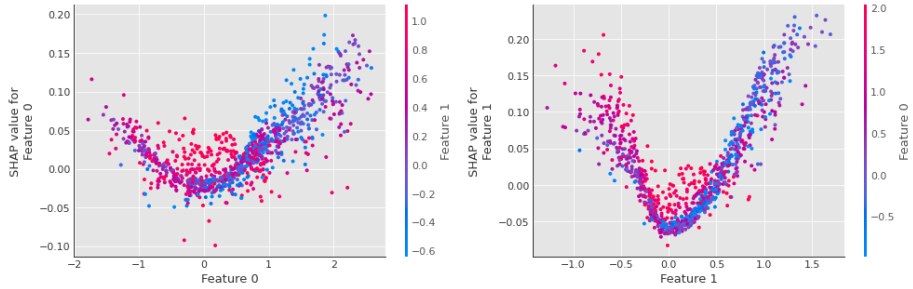
Figure A.2: Digits from the MNIST dataset were altered by striking them through, to create out-of-distribution data. The top row shows the unchanged MNIST data, while the bottom row shows the corresponding stricken through MNIST data.

Epistemic uncertainty explanation

We explained the contribution of each pixel to the estimated epistemic uncertainty (A.2) with SHAP. As epistemic uncertainty is a proxy for out-of-distributionness, we first performed an experiment where we made our own out-of-distribution data for the MNIST dataset. We built a dataset in which we added noise by striking through each digit in the test set, as shown in Figure A.2. We then showed how the SHAP values detected this strike-through as being a source of epistemic uncertainty. As a second experiment, we visualized the sources of epistemic uncertainty in the unaltered MNIST data. These data are not out-of-distribution, but some are more atypical than others. We picked the 5% samples with the highest epistemic uncertainty for every class, and visualized which pixels were responsible for the high epistemic uncertainty in each of these examples.

Aleatoric uncertainty explanation

We also explained the contribution of each pixel to the estimated aleatoric uncertainty (A.1) with SHAP. First, we visualized the learned features of the neural network with their aleatoric uncertainty, using the UMAP algorithm [105], to verify if aleatoric uncertainty was maximal at the class boundaries. Then, we visualized what pixels were responsible for these confusions at the class boundaries. For five of the most common confusions between classes, we visualized the data for which the neural network made the wrong decision.We then visualized the explanations for those samples, revealing which pixels were responsible for the confusion. For all wrongly predicted samples, we also computed the Pearson correlation between the class predicted as the second most likely, and the SHAP value for the aleatoric uncertainty. We expect a



(a) (b)

Figure A.3: Dependence plots, each showing the effect of a single feature on the estimated epistemic uncertainty.

positive correlation, as the evidence for the second most likely class should be causing the confusion, and hence, the aleatoric uncertainty.

A.4 Results

A.4.1 Synthetic dataset: two moons

Epistemic uncertainty

Figure A.3 shows the two SHAP dependence plots for the epistemic uncertainty of the model. The influence of feature 0 or x_1 on the estimate of epistemic uncertainty is demonstrated in Figure A.3a and the influence of feature 1 or x_2 is demonstrated in Figure A.3b. The y-axis shows the SHAP values for each feature, respectively, which represent how knowing that feature influences the estimated epistemic uncertainty of the model. This clearly demonstrates that the epistemic uncertainty is much lower inside the circle, and higher outside it. This is in line with expectations, as the epistemic uncertainty reflects what the model doesn't know, and is hence a proxy for 'out-of-distributionness'.

Figure A.4a shows the epistemic uncertainty for all the test set samples. Epistemic uncertainty clearly increases as we move away from the decision boundary, and from the circle representing the training data boundary. To see which features are responsible for this increased epistemic uncertainty, Figure A.4b demonstrates the SHAP values of x_1 and x_2. The green color indicates samples for which the SHAP values of x_1 are larger than those of x_2, and pink is used for the samples where the opposite is true. Checking the

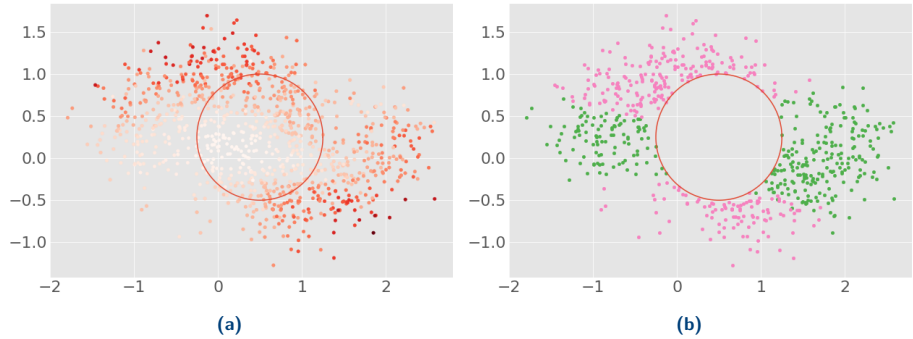


Figure A.4: Epistemic uncertainty visualizations. (a) Total estimated epistemic uncertainty. The color intensity represents the value of the epistemic uncertainty, demonstrating that epistemic uncertainty increases as we move away from the decision boundary, and from the circle representing the training data boundary. (b) The out-of-distribution samples, with a color indicating which of the two coordinates had a higher SHAP value, and thus a larger effect on the epistemic uncertainty.

SHAP values' ability to discern which feature was more out-of-distribution, by comparing which feature's SHAP value was larger, we got an 80.1% accuracy.

Aleatoric uncertainty

Figure A.5 shows the two SHAP dependence plots for the aleatoric uncertainty of the model. The influence of feature 0 or x_1 on the estimate of aleatoric uncertainty is demonstrated in Figure A.5a and the influence of feature 1 or x_2 is demonstrated in Figure A.5b. The y-axis shows how knowing a feature influences the estimated aleatoric uncertainty of the model. This clearly demonstrates that the aleatoric uncertainty higher in the middle of the distribution than at the sides. As the aleatoric uncertainty is inherent uncertainty due to noise and randomness, it reflects the uncertainty associated with imperfect classifications at the decision boundary. Therefore, aleatoric uncertainty is expected to be the highest at the decision boundary, and these scatter plots reflect this.

Figure A.6 shows a more detailed study of how the aleatoric uncertainty is influenced by both inputs. As shown in Figure A.6b, x_1 contributes most to the aleatoric uncertainty on the band at a 45°angle, corresponding to the main diagonal in the two moons dataset. Figure A.6d demonstrates that x_2 contributes more to the aleatoric uncertainty on the horizontal band in the middle of the moons corresponding to the main diagonal in the two moons dataset. We can best understand these explanations of uncertainty by looking at the SHAP explanations of the model predictions. x_1 is most informative for the

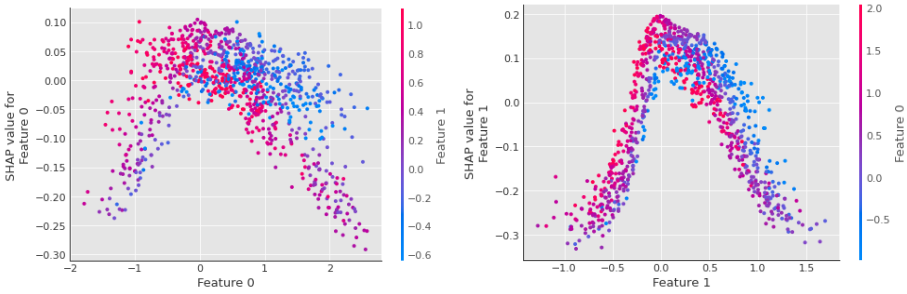


Figure A.5: Dependence plots, each showing the effect of a single feature on the estimated aleatoric uncertainty.

model prediction at the left-most and right-most side of the moons, as we don't need to know the value of x_2 there to make a correct prediction (see Figure A.1). x_2 is most informative for the model prediction at the top and bottom of the moons, as the prediction does not depend on x_1 there (see Figure A.1). We see that x_1 and x_2's contributions to the uncertainty, respectively, show the opposite pattern to x_1 and x_2's contributions to the predictions. The Pearson correlation between both explanations is (-0.37, p=1e-34) for x_1, and (-0.59, p=1e-96) for x_2.

A.4.2 MNIST dataset

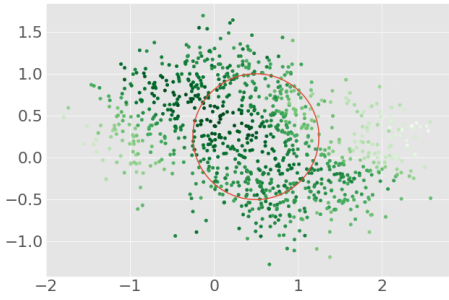
Epistemic uncertainty

We first demonstrated the ability of SHAP values to reveal that an added source of noise (a line which struck through the digits) was a source of epistemic uncertainty. Figure A.7 shows an image which is the average of all test set images, with the corresponding average SHAP value image revealing which pixels explained the epistemic uncertainty. The line to strike through the digits was always in the same location. It can clearly be discerned in both the average image and in the epistemic uncertainty SHAP value image. Importantly, the strike-through line is marked in red, which reveals a high average contribution to epistemic uncertainty. The Mann-Whitney U test for independent samples confirms that the values of the epistemic uncertainty SHAP values were significantly higher for the strike-through line than for the digits themselves, with (statistic= $2.61e11$, $p = 0.0$).

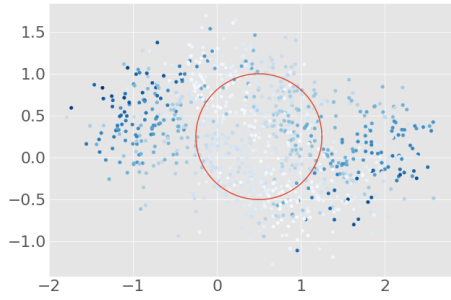
Second, we showed the sources of epistemic uncertainty in the unaltered MNIST



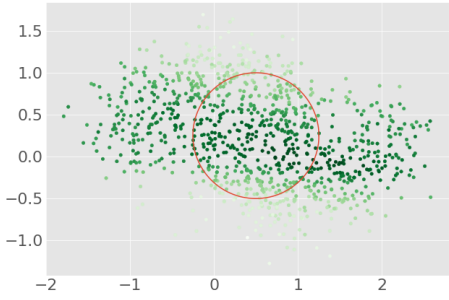
(a) The total aleatoric uncertainty



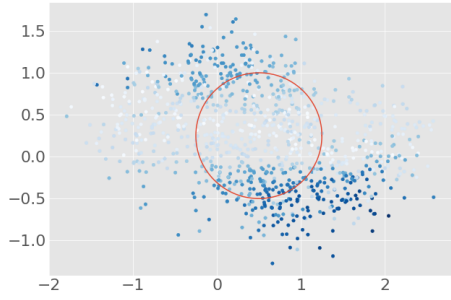
(b) SHAP values of aleatoric uncertainty for x_1



(c) SHAP values of model predictions for x_1



(d) SHAP values of aleatoric uncertainty for x_2



(e) SHAP values of model predictions for x_2

Figure A.6: Aleatoric uncertainty visualizations. (a) Total estimated aleatoric uncertainty. The color intensity represents the value of the aleatoric uncertainty, demonstrating that it is maximal nearest to the decision boundary. (b) and (c) show the SHAP values for feature x_1's contribution to aleatoric uncertainty, and to the model's prediction, respectively. There is a clear and significant negative correlation between the two. (d) and (e) show the same for x_2.

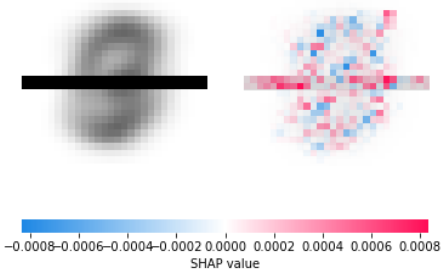


Figure A.7: A line that strikes through the MNIST dataset's digits is detected as a source of epistemic uncertainty. Left: the average of all struck-through test set digits of the MNIST dataset. Note that the reason why the strike-through line is darker than the rest of the image, is the fact that it is always in the same location. Right: the average of the pixels' SHAP values explaining the epistemic uncertainty. We see that the strike-through line is marked in red, revealing a high average contribution to epistemic uncertainty.

dataset. and its sources of epistemic uncertainty visualized. Figure A.8a shows the average digits of the MNIST test dataset, with the SHAP values of the model's prediction. Figure A.8b shows the 5% data with the highest epistemic uncertainty in the MNIST test dataset. Compared to the data shown in Figure A.8a, these data show more variability, with a more fuzzy average image. As the data are variable, so are the sources of epistemic uncertainty, shown in the SHAP value plots on their right. The pixels responsible for the epistemic uncertainty are largely found within the digits, mainly at locations that do not correspond to the average 'normal' digit. For example, the cross stroke through the seven is a rare feature, and thus a source of epistemic uncertainty, as is the serif extending from the base of the one.

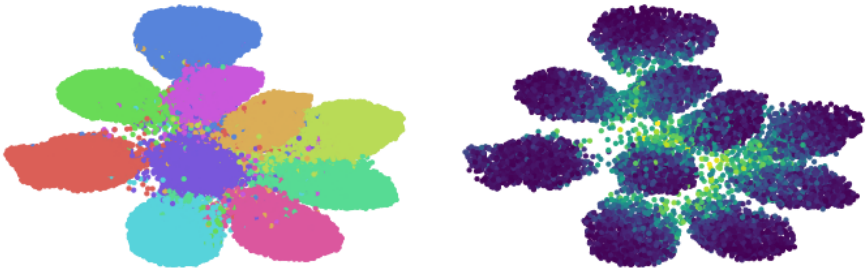
Aleatoric uncertainty

First, we visualize the output of the last hidden layer of the neural network architecture in Figure A.9, using two different colorings. Figure A.9a shows the distribution of the 10 MNIST classes in this layer. The aleatoric uncertainty coloring in Figure A.9b provides an insight into the global reasons for aleatoric uncertainty: this uncertainty stems from confusion at the class boundaries.

For the 438 images that were wrongly classified in the test set, the aleatoric uncertainty was significantly correlated with the prototype of the second most likely digit. The Pearson correlation was significant, ($\rho = 0.00767$, $p = 7.01e-6$). For the same set of wrongly predicted images, we then investigated some of the



Figure A.8: The MNIST dataset and its sources of epistemic uncertainty visualized. (a) The average of every digit in the test set, grouped by label (0 to 9). On the right of each digit are the SHAP values of the model's prediction, corresponding to the true class. (b) The average of the 5% images with the highest epistemic uncertainty. On the right are the SHAP values of the epistemic uncertainty for the same images. This reveals that the sources of epistemic uncertainty are quite variable across the MNIST dataset.



(a) (b)

Figure A.9: UMAP visualization showing the output of the last hidden layer of the neural network architecture trained on the MNIST dataset. This demonstrates how the 10 classes are internally represented in the network in the layer before the classification layer. The whole test set is represented, so each dot corresponds to one MNIST image of the test set. The UMAP is shown with two different colorings, where each sample is colored according to (a) its class, (b) the aleatoric uncertainty of the sample, with brighter colors corresponding to higher uncertainty.

most common confusions: 9 wrongly classified as 4, 5 as 3, 9 as 7, 3 as 8 and 7 as 2. Figure A.10 explains these confusions between the classes through images and their corresponding SHAP values. For this image, we averaged the images and SHAP values over all instances for which these particular confusions happened (varying from 14 to 26 instances). Additionally to showing the average images themselves, this figure shows the SHAP values for the model predictions with respect to all 10 classes. Lastly, it visualizes the SHAP values of the aleatoric uncertainty. These clearly show features of the two classes which are responsible for the confusion, highlighting features of the true class and/or the predicted class. Hence, the aleatoric uncertainty SHAP values show similarities with the SHAP values of the true class, the SHAP values of the the predicted class or both.

A.5 Discussion

To our knowledge, this is the first work that tackles explaining uncertainty with post-hoc explainability methods with distinguishing between aleatoric and epistemic uncertainty.

An important first topic of discussion is the relevance of explaining the uncertainty: is this really useful, and if so, is it mainly useful for developers or for users of AI? When it comes to epistemic uncertainty, the usefulness of

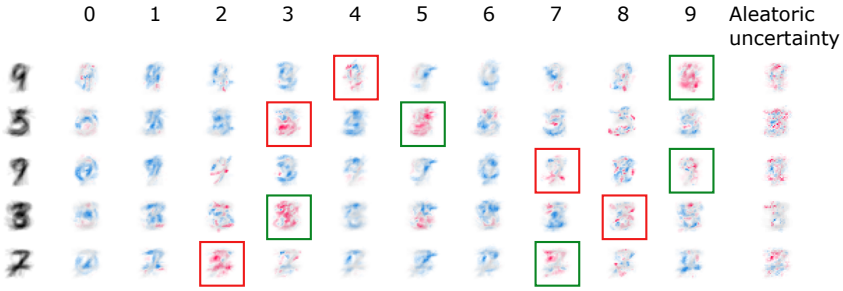


Figure A.10: Explanations for five of the most common confusions between digits. The first column shows the average MNIST image for these confused digits in the test set, the 'prototypes' of these confused digits. The ten following columns show the average SHAP values of the network predictions for the ten classes in MNIST. The red box shows the wrongly predicted digit, and the green box indicates the correct class. The average SHAP values for the aleatoric uncertainty are shown in the last column. This shows features that are common between the confused classes, or that caused the confusion.

explanations is very easy to grasp. Indeed, as shown in Figure A.7, explaining epistemic uncertainty can help to detect sources of noise that lead to out-of-distribution data. In our experiments, this was shown with a strike through MNIST digits: explaining epistemic uncertainty helps to identify that stripe.

The advantage of aleatoric uncertainty explanations is less straight-forward: we could also show explanations for the different classes to convey the same information. However, Figure A.10 shows us a potential advantage of aleatoric uncertainty explanations opposed to explanations for the individual classes. With aleatoric uncertainty explanations, we can show the user one image to explain a model's confusion or uncertainty. This is in contrast to having to provide multiple images, with the SHAP values for the different classes that it is considering. With aleatoric uncertainty explanations, the user may receive a simple message like "The neural network hesitates between 4 and 9, and the reason is the following:" followed by an image of the aleatoric uncertainty explanation.

In conclusion, uncertainty explanation can be a valuable tool for identifying sources of machine learning uncertainty. By leveraging uncertainty estimation methods that distinguish between epistemic and aleatoric uncertainty, SHAP values can pinpoint the exact causes of these uncertainties for the user. In future work, we plan to apply this method to sleep staging, aiming to enhance its utility in clinical practice.

Bibliography

[1] L. Besedovsky, T. Lange, and M. Haack, "The Sleep-Immune Crosstalk in Health and Disease," *Physiological Reviews*, vol. 99, no. 3, p. 1325, jul 2019. [Online]. Available: /pmc/articles/PMC6689741//pmc/articles/PMC6689741/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC6689741/

[2] E. Underwood, "Sleep: The brain's housekeeper?" *Science*, vol. 342, no. 6156, pp. 301–301, 2013. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.342.6156.301

[3] W. D. Killgore, "Effects of sleep deprivation on cognition," ser. Progress in Brain Research, G. A. Kerkhof and H. P. van Dongen, Eds. Elsevier, 2010, vol. 185, pp. 105–129. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780444537027000075

[4] S. Jakowski, A. Kiel, L. Kullik, and D. Erlacher, "Sleep to heal and restore: The role of sleep in the recovery and regeneration process," in *The Importance of Recovery for Physical and Mental Health: Negotiating the Effects of Underrecovery*. Taylor and Francis, jan 2023, pp. 141–153. [Online]. Available: https://www.taylorfrancis.com/chapters/edit/10.4324/9781003250647-9/sleep-heal-restore-sarah-jakowski-asja-kiel-lisa-kullik-daniel-erlacher

[5] S. Taheri, "The link between short sleep duration and obesity: we should recommend more sleep to prevent obesity," *Archives of Disease in Childhood*, vol. 91, no. 11, p. 881, nov 2006. [Online]. Available: /pmc/articles/PMC2082964//pmc/articles/PMC2082964/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC2082964/

[6] K. Wulff, S. Gatti, J. G. Wettstein, and R. G. Foster, "Sleep and circadian rhythm disruption in psychiatric and neurodegenerative disease," *Nature Reviews Neuroscience*, vol. 11, no. 8, pp. 589–599, 2010. [Online]. Available: https://doi.org/10.1038/nrn2868

[7] R. Berry, S. Quan, and A. Abreu, *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications, Version 2.6*. Darien: American Academy of Sleep Medicine, 2020.

[8] A. K. Patel, V. Reddy, K. R. Shumway, and J. F. Araujo, "Physiology, sleep stages," *StatPearls*, Jan 2024, copyright © 2024, StatPearls Publishing LLC. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK526132/

[9] D. Dawson and K. Reid, "Fatigue, alcohol and performance impairment," *Nature*, vol. 388, no. 6639, p. 235, 1997. [Online]. Available: https://doi.org/10.1038/40775

[10] J. Pan, J. Wu, J. Liu, J. Wu, and F. Wang, "A systematic review of sleep in patients with disorders of consciousness: From diagnosis to prognosis," *Brain Sciences*, vol. 11, no. 8, aug 2021.

[11] Institute of Medicine, *Sleep Disorders and Sleep Deprivation: An Unmet Public Health Problem*. Washington, DC: The National Academies Press, 2006.

[12] H. Di, Y. Guo, I. Daghlas, L. Wang, G. Liu, A. Pan, L. Liu, and Z. Shan, "Evaluation of Sleep Habits and Disturbances Among US Adults, 2017-2020," *JAMA Network Open*, vol. 5, no. 11, pp. e2 240 788–e2 240 788, nov 2022. [Online]. Available: https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2798209

[13] RAND Europe, "Why sleep matters: Quantifying the economic costs of insufficient sleep," https://www.rand.org/randeurope/research/projects/the-value-of-the-sleep-economy.html, 2023, accessed: 2024-04-22.

[14] S. Schutte-Rodin, L. Broch, D. Buysse, C. Dorsey, and M. Sateia, "Clinical guideline for the evaluation and management of chronic insomnia in adults," *Journal of Clinical Sleep Medicine*, vol. 04, no. 05, pp. 487–504, 2008. [Online]. Available: https://jcsm.aasm.org/doi/abs/10.5664/jcsm.27286

[15] K. A. Franklin and E. Lindberg, "Obstructive sleep apnea is a common disorder in the population—a review on the epidemiology of sleep apnea," *Journal of Thoracic Disease*, vol. 7, no. 8, p. 1311, 2015. [Online]. Available: /pmc/articles/PMC4561280//pmc/articles/PMC4561280/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4561280/

[16] J. Van Erum, D. Van Dam, and P. P. De Deyn, "Alzheimer's disease: Neurotransmitters of the sleep-wake cycle," pp. 72–80, oct 2019.

[17] M. A. Lloret, A. Cervera-Ferri, M. Nepomuceno, P. Monllor, D. Esteve, and A. Lloret, "Is Sleep Disruption a Cause or Consequence of Alzheimer's Disease? Reviewing Its Possible Role as a Biomarker," *International Journal of Molecular Sciences*, vol. 21, no. 3, feb 2020. [Online]. Available: /pmc/articles/PMC7037733//pmc/articles/PMC7037733/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC7037733/

[18] B. Guarnieri, F. Adorni, M. Musicco, I. Appollonio, E. Bonanni, P. Caffarra, C. Caltagirone, G. Cerroni, L. Concari, F. I. Cosentino, S. Ferrara, S. Fermi, R. Ferri, G. Gelosa, G. Lombardi, D. Mazzei, S. Mearelli, E. Morrone, L. Murri, F. M. Nobili, S. Passero, R. Perri, R. Rocchi, P. Sucapane, G. Tognoni, S. Zabberoni, and S. Sorbi, "Prevalence of sleep disturbances in mild cognitive impairment and dementing disorders: a multicenter Italian clinical cross-sectional study on 431 patients," *Dementia and geriatric cognitive disorders*, vol. 33, no. 1, pp. 50–58, mar 2012. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/22415141/

[19] M. R. Klauber, N. Butters, L. Parker, and D. F. Kripke, "Dementia in Institutionalized Elderly: Relation to Sleep Apnea," *Journal of the American Geriatrics Society*, vol. 39, no. 3, pp. 258–263, mar 1991. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1111/j.1532-5415.1991.tb01647.xhttps://onlinelibrary.wiley.com/doi/abs/10.1111/j.1532-5415.1991.tb01647.xhttps://agsjournals.onlinelibrary.wiley.com/doi/10.1111/j.1532-5415.1991.tb01647.x

[20] D. W. Kang, C. U. Lee, and H. K. Lim, "Role of Sleep Disturbance in the Trajectory of Alzheimer's Disease," *Clinical Psychopharmacology and Neuroscience*, vol. 15, no. 2, p. 89, 2017. [Online]. Available: /pmc/articles/PMC5426492//pmc/articles/PMC5426492/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC5426492/

[21] R. Fronczek, S. van Geest, M. Frölich, S. Overeem, F. W. Roelandse, G. J. Lammers, and D. F. Swaab, "Hypocretin (orexin) loss in Alzheimer's disease," *Neurobiology of Aging*, vol. 33, no. 8, pp. 1642–1650, aug 2012.

[22] J. E. Kang, M. M. Lim, R. J. Bateman, J. J. Lee, L. P. Smyth, J. R. Cirrito, N. Fujiki, S. Nishino, and D. M. Holtzman, "Amyloid-β dynamics are regulated by orexin and the sleep-wake cycle," *Science*, vol. 326, no. 5955, pp. 1005–1007, nov 2009. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.1180962

[23] L. Shi, S. J. Chen, M. Y. Ma, Y. P. Bao, Y. Han, Y. M. Wang, J. Shi, M. V. Vitiello, and L. Lu, "Sleep disturbances increase the risk of dementia: A systematic review and meta-analysis," *Sleep*

medicine reviews, vol. 40, pp. 4–16, aug 2018. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/28890168/

[24] J. E. Gaugler, A. B. Edwards, E. E. Femia, S. H. Zarit, M. A. P. Stephens, A. Townsend, and R. Greene, "Predictors of Institutionalization of Cognitively Impaired Elderbut Family Help and the Timing of Placement," *The Journals of Gerontology: Series B*, vol. 55, no. 4, pp. P247–P255, jul 2000. [Online]. Available: https://academic.oup.com/psychsocgerontology/article/55/4/P247/572857

[25] C. P. Pollak and D. Perlick, "Sleep Problems and Institutionalization of the Elderly," *https://doi.org/10.1177/089198879100400405*, vol. 4, no. 4, pp. 204–210, oct 1991. [Online]. Available: https://journals.sagepub.com/doi/10.1177/089198879100400405?url{_}ver=Z39.88-2003{&}rfr{_}id=ori{%}3Arid{%}3Acrossref.org{&}rfr{_}dat=cr{_}pub++0pubmed

[26] J. V. Rundo and R. Downey, "Polysomnography," in *Handbook of Clinical Neurology*. Elsevier B.V., 1 2019, vol. 160, pp. 381–392.

[27] D. Testelmans, "personal communication," mar 2020.

[28] A. Crivello, P. Barsocchi, M. Girolami, and F. Palumbo, "The meaning of sleep quality: A survey of available technologies," *IEEE Access*, vol. 7, pp. 167 374–167 390, 2019.

[29] S. A. Imtiaz, "A systematic review of sensing technologies for wearable sleep staging," *Sensors*, vol. 21, no. 5, 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/5/1562

[30] S. Debener, R. Emkes, M. De Vos, and M. Bleichner, "Unobtrusive ambulatory EEG using a smartphone and flexible printed electrodes around the ear," *Scientific Reports*, vol. 5, no. 1, pp. 1–11, nov 2015.

[31] "SeizeIT2 - EIT Health," accessed on: [insert date here]. [Online]. Available: https://eithealth.eu/project/seizeit2/

[32] K. B. Mikkelsen, Y. R. Tabar, S. L. Kappel, C. B. Christensen, H. O. Toft, M. C. Hemmsen, M. L. Rank, M. Otto, and P. Kidmose, "Accurate whole-night sleep monitoring with dry-contact ear-EEG," *Scientific Reports*, vol. 9, no. 1, pp. 1–12, dec 2019.

[33] S. L. Kappel, M. L. Rank, H. O. Toft, M. Andersen, and P. Kidmose, "Dry-Contact Electrode Ear-EEG," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 1, pp. 150–158, 1 2019.

[34] C. T. Lin, C. H. Chuang, Z. Cao, A. K. Singh, C. S. Hung, Y. H. Yu, M. Nascimben, Y. T. Liu, J. T. King, T. P. Su, and S. J. Wang, "Forehead EEG in Support of Future Feasible Personal Healthcare Solutions: Sleep Management, Headache Prevention, and Depression Treatment," *IEEE Access*, vol. 5, pp. 10 612–10 621, 2017.

[35] K. B. Mikkelsen, J. K. Ebajemito, M. A. Bonmati-Carrion, N. Santhi, V. L. Revell, G. Atzori, C. della Monica, S. Debener, D. Dijk, A. Sterr, and M. De Vos, "Machine-learning-derived sleep–wake staging from around-the-ear electroencephalogram outperforms manual scoring and actigraphy," *Journal of Sleep Research*, vol. 28, no. 2, 4 2019. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/jsr.12786

[36] P. J. Arnal, V. Thorey, E. Debellemaniere, M. E. Ballard, A. Bou Hernandez, A. Guillot, H. Jourde, M. Harris, M. Guillard, P. Van Beers, M. Chennaoui, and F. Sauvet, "The Dreem Headband compared to polysomnography for electroencephalographic signal acquisition and sleep staging," *Sleep*, vol. 43, no. 11, pp. 1–13, nov 2020. [Online]. Available: https://academic.oup.com/sleep/article/43/11/zsaa097/5841249

[37] C. S. Nayak and A. C. Anilkumar, "EEG Normal Sleep," *StatPearls*, may 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK537023/

[38] A. F. Agarap, "Deep learning using rectified linear units (relu)," *arXiv preprint arXiv:1803.08375*, 2018.

[39] L. Fiorillo, A. Puiatti, M. Papandrea, P. L. Ratti, P. Favaro, C. Roth, P. Bargiotas, C. L. Bassetti, and F. D. Faraci, "Automated sleep scoring: A review of the latest approaches," *Sleep Medicine Reviews*, vol. 48, p. 101204, dec 2019.

[40] P. Chriskos, C. A. Frantzidis, C. M. Nday, P. T. Gkivogkli, P. D. Bamidis, and C. Kourtidou-Papadeli, "A review on current trends in automatic sleep staging through bio-signal recordings and future challenges," *Sleep Medicine Reviews*, vol. 55, p. 101377, feb 2021.

[41] H. Phan and K. Mikkelsen, "Automatic sleep staging of EEG signals: recent development, challenges, and future directions," *Physiological Measurement*, vol. 43, no. 4, p. 04TR01, apr 2022. [Online]. Available: https://iopscience.iop.org/article/10.1088/1361-6579/ac6049

[42] M. Gaiduk, Á. Serrano Alarcón, R. Seepold, and N. Martínez Madrid, "Current status and prospects of automatic sleep stages scoring: Review," *Biomedical Engineering Letters*,

vol. 13, no. 3, p. 247, aug 2023. [Online]. Available: /pmc/articles/PMC10382458//pmc/articles/PMC10382458/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC10382458/

[43] M. Perslev, S. Darkner, L. Kempfner, M. Nikolic, P. J. Jennum, and C. Igel, "U-Sleep: resilient high-frequency sleep staging," *npj Digital Medicine*, vol. 4, no. 1, p. 72, dec 2021. [Online]. Available: http://www.nature.com/articles/s41746-021-00440-5

[44] S. Chambon, M. N. Galtier, P. J. Arnal, G. Wainrib, and A. Gramfort, "A Deep Learning Architecture for Temporal Sleep Stage Classification Using Multivariate and Multimodal Time Series," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 4, pp. 758–769, apr 2018.

[45] H. Sun, J. Jia, B. Goparaju, G. B. Huang, O. Sourina, M. T. Bianchi, and M. B. Westover, "Large-Scale Automated Sleep Staging," *Sleep*, vol. 40, no. 10, oct 2017. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/29029305/

[46] H. Phan, F. Andreotti, N. Cooray, O. Y. Chén, and M. De Vos, "SeqSleepNet: End-to-End Hierarchical Recurrent Neural Network for Sequence-to-Sequence Automatic Sleep Staging," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 3, pp. 400–410, sep 2018. [Online]. Available: http://dx.doi.org/10.1109/TNSRE.2019.2896659

[47] A. Guillot and V. Thorey, "RobustSleepNet: Transfer Learning for Automated Sleep Staging at Scale," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 1441–1451, 2021.

[48] H. Phan, O. Y. Chén, M. C. Tran, P. Koch, A. Mertins, and M. De Vos, "XSleepNet: Multi-View Sequential Model for Automatic Sleep Staging," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, jul 2020. [Online]. Available: http://arxiv.org/abs/2007.05492http://dx.doi.org/10.1109/TPAMI.2021.3070057

[49] J. Fan, C. Sun, M. Long, C. Chen, and W. Chen, "EOGNET: A Novel Deep Learning Model for Sleep Stage Classification Based on Single-Channel EOG Signal," *Frontiers in Neuroscience*, vol. 15, p. 573194, jul 2021. [Online]. Available: www.frontiersin.org

[50] H. Phan, K. Mikkelsen, O. Y. Chén, P. Koch, A. Mertins, and M. De Vos, "SleepTransformer: Automatic Sleep Staging with Interpretability and Uncertainty Quantification," *IEEE Transactions on Biomedical*

Engineering, vol. 69, no. 8, pp. 2456–2467, may 2021. [Online]. Available: http://arxiv.org/abs/2105.11043http://dx.doi.org/10.1109/TBME.2022.3147187

[51] E. Hartmann, "The 90-Minute Sleep-Dream Cycle," *Archives of General Psychiatry*, vol. 18, no. 3, pp. 280–286, 03 1968. [Online]. Available: https://doi.org/10.1001/archpsyc.1968.01740030024004

[52] Z. Jia, Y. Lin, J. Wang, X. Wang, P. Xie, and Y. Zhang, "SalientSleepNet: Multimodal Salient Wave Detection Network for Sleep Staging," *IJCAI International Joint Conference on Artificial Intelligence*, pp. 2614–2620, may 2021. [Online]. Available: https://arxiv.org/abs/2105.13864v1

[53] Z. Jia, Y. Lin, J. Wang, R. Zhou, X. Ning, Y. He, and Y. Zhao, "GraphSleepNet: Adaptive Spatial-Temporal Graph Convolutional Networks for Sleep Stage Classification," *IJCAI International Joint Conference on Artificial Intelligence*, vol. 2, pp. 1324–1330, jul 2020. [Online]. Available: https://www.ijcai.org/proceedings/2020/184

[54] E. Eldele, Z. Chen, C. Liu, M. Wu, C. K. Kwoh, X. Li, and C. Guan, "An Attention-Based Deep Learning Approach for Sleep Stage Classification with Single-Channel EEG," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 809–818, 2021.

[55] H. Seo, S. Back, S. Lee, D. Park, T. Kim, and K. Lee, "Intra- and inter-epoch temporal context network (IITNet) using sub-epoch features for automatic sleep scoring on raw single-channel EEG," *Biomedical Signal Processing and Control*, vol. 61, p. 102037, aug 2020.

[56] A. Olesen, P. Jennum, E. Mignot, and H. . Sorensen, "Automatic sleep stage classification with deep residual networks in a mixed-cohort setting," 2020.

[57] H. van Gorp, I. A. M Huijben, P. Fonseca, R. J. G van Sloun, S. Overeem, M. M. van Gilst, and C. author Hans van Gorp, "Certainty about uncertainty in sleep staging: a theoretical framework," *Sleep*, vol. 2022, pp. 1–8, jun 2022. [Online]. Available: https://academic.oup.com/sleep/advance-article/doi/10.1093/sleep/zsac134/6604464

[58] R. S. Rosenberg and S. V. Hout, "The American Academy of Sleep Medicine Inter-scorer Reliability Program: Sleep Stage Scoring," *Journal of Clinical Sleep Medicine : JCSM : Official Publication of the American Academy of Sleep Medicine*, vol. 9, no. 1, p. 81, 2013. [Online]. Available: /pmc/articles/PMC3525994//pmc/articles/PMC3525994/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3525994/

[59] Y. J. Lee, . Jae, Y. Lee, H. Cho, . Ji, and H. Choi, "Interrater reliability of sleep stage scoring: a meta-analysis," *J Clin Sleep Med*, vol. 18, pp. 193–202, 2022. [Online]. Available: https://doi.org/10.5664/jcsm.9538

[60] X. Zhang, X. Dong, J. W. Kantelhardt, J. Li, L. Zhao, C. Garcia, M. Glos, T. Penzel, and F. Han, "Process and outcome for international reliability in sleep scoring," *Sleep & breathing = Schlaf & Atmung*, vol. 19, no. 1, pp. 191–195, mar 2015. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/24801137/

[61] A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun, P. Sundberg, H. Yee, K. Zhang, Y. Zhang, G. Flores, G. E. Duggan, J. Irvine, Q. Le, K. Litsch, A. Mossin, J. Tansuwan, D. Wang, J. Wexler, J. Wilson, D. Ludwig, S. L. Volchenboum, K. Chou, M. Pearson, S. Madabushi, N. H. Shah, A. J. Butte, M. D. Howell, C. Cui, G. S. Corrado, and J. Dean, "Scalable and accurate deep learning with electronic health records," *NPJ digital medicine*, vol. 1, no. 1, dec 2018. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/31304302/

[62] D. S. Ting, Y. Liu, P. Burlina, X. Xu, N. M. Bressler, and T. Y. Wong, "AI for medical imaging goes deep," *Nature Medicine*, vol. 24, no. 5, pp. 539–540, may 2018. [Online]. Available: https://www.nature.com/articles/s41591-018-0029-3

[63] J. Egger, C. Gsaxner, A. Pepe, K. L. Pomykala, F. Jonske, M. Kurz, J. Li, and J. Kleesiek, "Medical deep learning—A systematic meta-review," *Computer Methods and Programs in Biomedicine*, vol. 221, p. 106874, jun 2022.

[64] M. Javanmardi and T. Tasdizen, "Domain adaptation for biomedical image segmentation using adversarial training," in Proceedings of the *Proceedings - International Symposium on Biomedical Imaging*, vol. 2018-April, pp. 554–558, may 2018.

[65] C. O'Reilly, N. Gosselin, J. Carrier, and T. Nielsen, "Montreal archive of sleep studies: An open-access resource for instrument benchmarking and exploratory research," *Journal of Sleep Research*, vol. 23, no. 6, pp. 628–635, 12 2014. [Online]. Available: http://physionet.org/,

[66] A. Kales and A. Rechtschaffen, *A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects*. Washington DC: United States Government Printing Office, 1968.

[67] C. Iber, S. Ancoli-Israel, A. L. Chesson, and S. Quan, "The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and

Technical Specifications," *Westchester, IL: American Academy of Sleep Medicine*, 2007.

[68] A. Sterr, J. K. Ebajemito, K. B. Mikkelsen, M. A. Bonmati-Carrion, N. Santhi, C. della Monica, L. Grainger, G. Atzori, V. Revell, S. Debener, D.-J. Dijk, and M. De Vos, "Sleep EEG Derived From Behind-the-Ear Electrodes (cEEGrid) Compared to Standard Polysomnography: A Proof of Concept Study," *Frontiers in Human Neuroscience*, vol. 12, p. 452, 11 2018. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnhum.2018.00452/full

[69] S. Debener, F. Minow, R. Emkes, K. Gandras, and M. de Vos, "How about taking a low-cost, small, and wireless EEG for a walk?" *Psychophysiology*, vol. 49, no. 11, pp. 1617–1621, 11 2012. [Online]. Available: http://doi.wiley.com/10.1111/j.1469-8986.2012.01471.x

[70] C. H. Bastien, A. Vallières, and C. M. Morin, "Validation of the Insomnia Severity Index as an outcome measure for insomnia research," *Sleep medicine*, vol. 2, no. 4, pp. 297–307, 2001. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/11438246/

[71] T. Becker, K. Vandecasteele, C. Chatzichristos, W. Van Paesschen, D. Valkenborg, S. Van Huffel, and M. De Vos, "Classification with a deferral option and low-trust filtering for automated seizure detection," *Sensors*, vol. 21, no. 4, pp. 1–18, 2 2021. [Online]. Available: https://doi.org/10.3390/s21041046

[72] K. Vandecasteele, T. De Cooman, J. Dan, E. Cleeren, S. Van Huffel, B. Hunyadi, and W. Van Paesschen, "Visual seizure annotation and automated seizure detection using behind-the-ear electroencephalographic channels," *Epilepsia*, vol. 61, no. 4, pp. 766–775, 4 2020.

[73] A. Devulder, J. Macea, A. Kalkanis, F.-L. De Winter, M. Vandenbulcke, R. Vandenberghe, D. Testelmans, M. J. A. Van Den Bossche, and W. Van Paesschen, "Subclinical epileptiform activity and sleep disturbances in alzheimer's disease," *Brain and Behavior*, vol. 13, no. 12, p. e3306, 2023. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/brb3.3306

[74] G. M. McKhann, D. S. Knopman, H. Chertkow, B. T. Hyman, C. R. Jack Jr., C. H. Kawas, W. E. Klunk, W. J. Koroshetz, J. J. Manly, R. Mayeux, R. C. Mohs, J. C. Morris, M. N. Rossor, P. Scheltens, M. C. Carrillo, B. Thies, S. Weintraub, and C. H. Phelps, "The diagnosis of dementia due to alzheimer's disease: Recommendations from the national institute on aging-alzheimer's association workgroups on

diagnostic guidelines for alzheimer's disease," *Alzheimer's & Dementia*, vol. 7, no. 3, pp. 263–269, 2011. [Online]. Available: https://alz-journals.onlinelibrary.wiley.com/doi/abs/10.1016/j.jalz.2011.03.005

[75] M. S. Albert, S. T. DeKosky, D. Dickson, B. Dubois, H. H. Feldman, N. C. Fox, A. Gamst, D. M. Holtzman, W. J. Jagust, R. C. Petersen, P. J. Snyder, M. C. Carrillo, B. Thies, and C. H. Phelps, "The diagnosis of mild cognitive impairment due to alzheimer's disease: Recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease," *Alzheimer's & Dementia*, vol. 7, no. 3, pp. 270–279, 2011. [Online]. Available: https://alz-journals.onlinelibrary.wiley.com/doi/abs/10.1016/j.jalz.2011.03.008

[76] M. F. Folstein, S. E. Folstein, and P. R. McHugh, ""mini-mental state": A practical method for grading the cognitive state of patients for the clinician," *Journal of Psychiatric Research*, vol. 12, no. 3, pp. 189–198, 1975. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0022395675900266

[77] H.-C. Huang, Y.-M. Tseng, Y.-C. Chen, P.-Y. Chen, and H.-Y. Chiu, "Diagnostic accuracy of the clinical dementia rating scale for detecting mild cognitive impairment and dementia: A bivariate meta-analysis," *International Journal of Geriatric Psychiatry*, vol. 36, no. 2, pp. 239–251, 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/gps.5436

[78] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," 2014.

[79] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: A System for Large-Scale Machine Learning," in Proceedings of the *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*, ser. OSDI'16, p. 265–283, 2016.

[80] H. Phan, O. Y. Chén, P. Koch, A. Mertins, and M. De Vos, "Deep transfer learning for single-channel automatic sleep staging with channel mismatch," in Proceedings of the *European Signal Processing Conference*, vol. 2019-Septe, 9 2019.

[81] H. Phan, O. Y. Chén, P. Koch, Z. Lu, I. Mcloughlin, A. Mertins, and M. De Vos, "Towards More Accurate Automatic Sleep Staging via Deep

Transfer Learning," *IEEE Transactions on Biomedical Engineering*, vol. 68, no. 6, pp. 1787–1798, 2021.

[82] H. Phan, K. Mikkelsen, O. Y. Chén, P. Koch, A. Mertins, P. Kidmose, and M. D. Vos, "Personalized automatic sleep staging with single-night data: a pilot study with Kullback–Leibler divergence regularization," *Physiological Measurement*, vol. 41, no. 6, p. 064004, 6 2020. [Online]. Available: https://iopscience.iop.org/article/10.1088/1361-6579/ab921ehttps://iopscience.iop.org/article/10.1088/1361-6579/ab921e/meta

[83] G. Wilson and D. J. Cook, "A Survey of Unsupervised Deep Domain Adaptation," *ACM Transactions on Intelligent Systems and Technology*, vol. 11, no. 5, pp. 1–46, 12 2018. [Online]. Available: http://arxiv.org/abs/1812.02849

[84] L. Hedegaard, O. A. Sheikh-Omar, and A. Iosifidis, "Supervised Domain Adaptation: A Graph Embedding Perspective and a Rectified Experimental Protocol," apr 2020. [Online]. Available: http://arxiv.org/abs/2004.11262

[85] Z. Wang, B. Du, and Y. Guo, "Domain Adaptation with Neural Embedding Matching," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 7, pp. 2387–2397, 7 2020.

[86] S. Motiian, M. Piccirilli, D. A. Adjeroh, and G. Doretto, "Unified Deep Supervised Domain Adaptation and Generalization," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2017-Octob, pp. 5716–5726, 9 2017. [Online]. Available: http://arxiv.org/abs/1709.10190

[87] X. Xu, X. Zhou, R. Venkatesan, G. Swaminathan, and O. Majumder, "d-SNE: Domain Adaptation using Stochastic Neighborhood Embedding," in Proceedings of the *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2492–2501, 2019.

[88] H. Hotelling, *Relations Between Two Sets of Variates*. New York, NY: Springer New York, 1992, pp. 162–190. [Online]. Available: https://doi.org/10.1007/978-1-4612-4380-9_14

[89] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in Proceedings of the *Proceedings of the 30th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, S. Dasgupta and D. McAllester, Eds., vol. 28, no. 3, pp. 1247–1255, 17–19 Jun 2013. [Online]. Available: https://proceedings.mlr.press/v28/andrew13.html

[90] P. Dhillon, D. P. Foster, and L. Ungar, "Multi-view learning of word embeddings via cca," in Proceedings of the *Advances in Neural Information Processing Systems*, J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, Eds., vol. 24, 2011. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2011/file/6c4b761a28b734fe93831e3fb400ce87-Paper.pdf

[91] R. Udupa and M. M. Khapra, "Transliteration equivalence using canonical correlation analysis," in Proceedings of the *Advances in Information Retrieval*, C. Gurrin, Y. He, G. Kazai, U. Kruschwitz, S. Little, T. Roelleke, S. Rüger, and K. van Rijsbergen, Eds., pp. 75–86, 2010.

[92] S. Chandar, M. M. Khapra, H. Larochelle, and B. Ravindran, "Correlational Neural Networks," *Neural Computation*, vol. 28, no. 2, pp. 257–285, 02 2016. [Online]. Available: https://doi.org/10.1162/NECO_a_00801

[93] Y. Xu, X. Xiang, and M. Huang, "Task-driven common representation learning via bridge neural network," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 5573–5580, Jul. 2019. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/4499

[94] S. Biswal, J. Kulas, H. Sun, B. Goparaju, M. B. Westover, M. T. Bianchi, and J. Sun, "SLEEPNET: Automated Sleep Staging System via Deep Learning," 7 2017. [Online]. Available: http://arxiv.org/abs/1707.08262

[95] S. Nasiri and G. D. Clifford, "Attentive Adversarial Network for Large-Scale Sleep Staging," *Proceedings of Machine Learning Research*, vol. 126, pp. 1–21, 2020.

[96] E. Eldele, M. Ragab, Z. Chen, M. Wu, C. K. Kwoh, X. Li, and C. Guan, "ADAST: Attentive Cross-Domain EEG-Based Sleep Staging Framework With Iterative Self-Training," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 7, no. 1, pp. 210–221, feb 2023.

[97] C. Yoo, H. W. Lee, and J. Kang, "Transferring Structured Knowledge in Unsupervised Domain Adaptation of a Sleep Staging Network," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 3, pp. 1273–1284, aug 2022.

[98] R. Berry, R. Brooks, C. Gamaldo, S. Harding, R. Lloyd, S. Quan, M. Troester, and B. Vaughn, "AASM Scoring Manual Updates for 2017 (Version 2.4)," *Journal of clinical sleep medicine : JCSM : official publication of the American Academy of Sleep Medicine*, vol. 13, 2017.

[99] M. Younes, M. Younes, and E. Giannouli, "Accuracy of automatic polysomnography scoring using frontal electrodes," *Journal of Clinical Sleep Medicine*, vol. 12, no. 5, pp. 735–746, 2016.

[100] M. M. Rahman, M. I. H. Bhuiyan, and A. R. Hassan, "Sleep stage classification using single-channel EOG," *Computers in Biology and Medicine*, vol. 102, pp. 211–220, 11 2018.

[101] K. Weiss, T. M. Khoshgoftaar, and D. D. Wang, "A survey of transfer learning," *Journal of Big Data*, vol. 3, no. 1, p. 9, 12 2016. [Online]. Available: http://journalofbigdata.springeropen.com/articles/10.1186/s40537-016-0043-6

[102] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[103] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep Domain Confusion: Maximizing for Domain Invariance," 12 2014. [Online]. Available: https://arxiv.org/abs/1412.3474v1

[104] M. Long, Y. Cao, J. Wang, M. I. Jordan, and J. Edu, "Learning Transferable Features with Deep Adaptation Networks," in Proceedings of the *ICML'15: Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, pp. 97–105, 2015.

[105] L. McInnes, J. Healy, N. Saul, and L. Großberger, "UMAP: Uniform Manifold Approximation and Projection," *Journal of Open Source Software*, vol. 3, no. 29, p. 861, 9 2018. [Online]. Available: http://joss.theoj.org/papers/10.21105/joss.00861

[106] X. Hong, Q. Zheng, L. Liu, P. Chen, K. Ma, Z. Gao, and Y. Zheng, "Dynamic Joint Domain Adaptation Network for Motor Imagery Classification," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 556–565, 2021.

[107] G. Bao, N. Zhuang, L. Tong, B. Yan, J. Shu, L. Wang, Y. Zeng, and Z. Shen, "Two-Level Domain Adaptation Neural Network for EEG-Based Emotion Recognition," *Frontiers in Human Neuroscience*, vol. 0, p. 620, 1 2021.

[108] H. Zhao, Q. Zheng, K. Ma, H. Li, and Y. Zheng, "Deep Representation-Based Domain Adaptation for Nonstationary EEG Classification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 2, pp. 535–545, feb 2021.

[109] X. Chai, Q. Wang, Y. Zhao, X. Liu, O. Bai, and Y. Li, "Unsupervised domain adaptation techniques based on auto-encoder for non-stationary EEG-based emotion recognition," *Computers in Biology and Medicine*, vol. 79, pp. 205–214, 12 2016.

[110] J. Li, S. Qiu, C. Du, Y. Wang, and H. He, "Domain adaptation for eeg emotion recognition based on latent representation similarity," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 12, no. 2, pp. 344–353, 6 2020.

[111] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-Adversarial Training of Neural Networks," *The Journal of Machine Learning Research*, vol. 17, pp. 1–35, may 2016. [Online]. Available: http://arxiv.org/abs/1505.07818

[112] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial Discriminative Domain Adaptation," in Proceedings of the *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, pp. 7167–7176, feb 2017. [Online]. Available: http://arxiv.org/abs/1702.05464

[113] D. Huysmans, P. Borzée, D. Testelmans, B. Buyse, T. Willemen, S. Van Huffel, and C. Varon, "Evaluation of a Commercial Ballistocardiography Sensor for Sleep Apnea Screening and Sleep Monitoring," *Sensors*, vol. 19, no. 9, p. 2133, 5 2019. [Online]. Available: https://www.mdpi.com/1424-8220/19/9/2133/htmhttps://www.mdpi.com/1424-8220/19/9/2133

[114] J. Fan, H. Zhu, X. Jiang, L. Meng, C. Chen, C. Fu, H. Yu, C. Dai, and W. Chen, "Unsupervised Domain Adaptation by Statistics Alignment for Deep Sleep Staging Networks," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 205–216, 2022.

[115] S. Yang, Z. Xie, H. Peng, M. Xu, M. Sun, and P. Li, "Dataset pruning: Reducing training data by examining generalization influence," in Proceedings of the *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=4wZiAXD29TQ

[116] D. Montes de Oca Zapiain, M. A. Wood, N. Lubbers, C. Z. Pereyra, A. P. Thompson, and D. Perez, "Training data selection for accuracy and transferability of interatomic potentials," *npj Computational Materials 2022 8:1*, vol. 8, no. 1, pp. 1–9, sep 2022. [Online]. Available: https://www.nature.com/articles/s41524-022-00872-x

[117] Y. Kwon and J. Zou, "Beta Shapley: a Unified and Noise-reduced Data Valuation Framework for Machine Learning," *Conference on Artificial Intelligence and Statistics (AISTATS)*, oct 2022. [Online]. Available: https://arxiv.org/abs/2110.14049v2

[118] B. Settles, "Active Learning Literature Survey," 2009.

[119] V. L. Nguyen, M. H. Shaker, and E. Hüllermeier, "How to measure uncertainty in uncertainty sampling for active learning," *Machine Learning*, vol. 111, no. 1, pp. 89–122, jan 2022. [Online]. Available: https://link.springer.com/article/10.1007/s10994-021-06003-9

[120] H. Kilian, P. Lorenzo, V. der Plas Dries, M. Wannes, D. Jesse, K. Hendrickx, L. Perini, D. Van der Plas, W. Meert, and J. Davis, "Machine Learning with a Reject Option: A survey," jul 2021. [Online]. Available: https://arxiv.org/abs/2107.11277v1

[121] N. Norori, Q. Hu, F. M. Aellen, F. D. Faraci, and A. Tzovara, "Addressing bias in big data and AI for health care: A call for open science," *Patterns*, vol. 2, no. 10, p. 100347, oct 2021.

[122] S. Pathak, C. Lu, S. B. Nagaraj, M. van Putten, and C. Seifert, "STQS: Interpretable multi-modal Spatial-Temporal-seQuential model for automatic Sleep scoring," *Artificial Intelligence in Medicine*, vol. 114, p. 102038, 4 2021.

[123] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, V. Makarenkov, and S. Nahavandi, "A Review of Uncertainty Quantification in Deep Learning: Techniques, Applications and Challenges," *Information Fusion*, vol. 76, pp. 243–297, nov 2020. [Online]. Available: http://arxiv.org/abs/2011.06225http://dx.doi.org/10.1016/j.inffus.2021.05.008

[124] E. Hüllermeier and W. Waegeman, "Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods," *Machine Learning*, vol. 110, no. 3, pp. 457–506, mar 2021. [Online]. Available: https://link.springer.com/article/10.1007/s10994-021-05946-3

[125] A. Guillot, F. Sauvet, E. H. During, and V. Thorey, "Dreem Open Datasets: Multi-Scored Sleep Datasets to Compare Human and Automated Sleep Staging," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 9, pp. 1955–1965, sep 2020.

[126] J. B. Stephansen, A. N. Olesen, M. Olsen, A. Ambati, E. B. Leary, H. E. Moore, O. Carrillo, L. Lin, F. Han, H. Yan, Y. L. Sun,

Y. Dauvilliers, S. Scholz, L. Barateau, B. Hogl, A. Stefani, S. C. Hong, T. W. Kim, F. Pizza, G. Plazzi, S. Vandi, E. Antelmi, D. Perrin, S. T. Kuna, P. K. Schweitzer, C. Kushida, P. E. Peppard, H. B. Sorensen, P. Jennum, and E. Mignot, "Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy," *Nature Communications*, vol. 9, no. 1, dec 2018. [Online]. Available: /pmc/articles/PMC6283836//pmc/articles/PMC6283836/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC6283836/

[127] I. A. Huijben, L. W. Hermans, A. C. Rossi, S. Overeem, M. M. van Gilst, and R. J. van Sloun, "Interpretation and further development of the hypnodensity representation of sleep structure," *Physiological Measurement*, vol. 44, no. 1, p. 015002, jan 2023. [Online]. Available: https://iopscience.iop.org/article/10.1088/1361-6579/aca641https://iopscience.iop.org/article/10.1088/1361-6579/aca641/meta

[128] P. Anderer, M. Ross, A. Cerny, R. Vasko, E. Shaw, and P. Fonseca, "Overview of the hypnodensity approach to scoring sleep for polysomnography and home sleep testing," *Frontiers in Sleep*, vol. 2, p. 1163477, apr 2023.

[129] L. Fiorillo, D. Pedroncelli, V. Agostini, P. Favaro, and F. D. Faraci, "Multi-scored sleep databases: how to exploit the multiple-labels in automated sleep scoring," *Sleep*, vol. 46, no. 5, pp. 1–12, may 2023. [Online]. Available: https://dx.doi.org/10.1093/sleep/zsad028

[130] D. Y. Kang, P. N. DeYoung, J. Tantiongloc, T. P. Coleman, and R. L. Owens, "Statistical uncertainty quantification to augment clinical decision support: a first implementation in sleep medicine," *npj Digital Medicine 2021 4:1*, vol. 4, no. 1, pp. 1–9, sep 2021. [Online]. Available: https://www.nature.com/articles/s41746-021-00515-3

[131] M. MacAs, N. Grimova, V. Gerla, L. Lhotska, and E. Saifutdinova, "Active Learning for Semiautomatic Sleep Staging and Transitional EEG Segments," *Proceedings - 2018 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2018*, pp. 2621–2627, jan 2019.

[132] L. Fiorillo, P. Favaro, and F. D. Faraci, "DeepSleepNet-Lite: A Simplified Automatic Sleep Stage Scoring Model with Uncertainty Estimates," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 2076–2085, 2021.

[133] J. Sebek, H. Schaabova, and V. Krajca, "Active learning approach for eeg classification using neural networks: A review," *2019 7th E-Health and Bioengineering Conference, EHB 2019*, nov 2019.

[134] M. Macaš, N. Grimová, V. Gerla, and L. Lhotská, "Semi-Automated Sleep EEG Scoring with Active Learning and HMM-Based Deletion of Ambiguous Instances," *Proceedings 2019, Vol. 31, Page 46*, vol. 31, no. 1, p. 46, nov 2019. [Online]. Available: https://www.mdpi.com/2504-3900/31/1/46

[135] N. Grimova, M. Macas, and V. Gerla, "Active learning for semi-automated sleep scoring," in Proceedings of the *Precision Medicine Powered by pHealth and Connected Health*, N. Maglaveras, I. Chouvarda, and P. de Carvalho, Eds., pp. 139–143, 2018.

[136] S. Swayamdipta, R. Schwartz, N. Lourie, Y. Wang, H. Hajishirzi, N. A. Smith, and Y. Choi, "Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics," *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pp. 9275–9293, sep 2020. [Online]. Available: https://arxiv.org/abs/2009.10795v2

[137] N. Seedat, J. Crabbé, I. Bica, and M. van der Schaar, "Data-iq: Characterizing subgroups with heterogeneous outcomes in tabular data," *Advances in Neural Information Processing Systems*, vol. 35, pp. 23 660–23 674, 2022.

[138] Y. Geifman and R. El-Yaniv, "Selective Classification for Deep Neural Networks," *Advances in Neural Information Processing Systems*, vol. 2017-December, pp. 4879–4888, may 2017. [Online]. Available: https://arxiv.org/abs/1705.08500v2

[139] R. El-Yaniv and Y. Wiener, "On the Foundations of Noise-free Selective Classification," *Journal of Machine Learning Research*, vol. 11, no. 53, pp. 1605–1641, 2010. [Online]. Available: http://jmlr.org/papers/v11/el-yaniv10a.html

[140] A. Olmin, "On uncertainty quantification in neural networks: Ensemble distillation and weak supervision," Ph.D. dissertation, Linköping University Electronic Press, 2022.

[141] M. H. Shaker and E. Hüllermeier, "Aleatoric and Epistemic Uncertainty with Random Forests," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12080 LNCS, pp. 444–456, jan 2020. [Online]. Available: https://arxiv.org/abs/2001.00893v1

[142] L. Fiorillo, G. Monachino, J. van der Meer, M. Pesce, J. D. Warncke, M. H. Schmidt, C. L. Bassetti, A. Tzovara, P. Favaro, and F. D.

Faraci, "U-Sleep's resilience to AASM guidelines," *npj Digital Medicine 2023 6:1*, vol. 6, no. 1, pp. 1–9, mar 2023. [Online]. Available: https://www.nature.com/articles/s41746-023-00784-0

[143] M. Younes, M. Ostrowski, M. Soiferman, H. Younes, M. Younes, J. Raneri, and P. Hanly, "Odds Ratio Product of Sleep EEG as a Continuous Measure of Sleep State," *Sleep*, vol. 38, no. 4, pp. 641–654, apr 2015. [Online]. Available: https://academic.oup.com/sleep/article/38/4/641/2416969

[144] M. Cesari, A. Stefani, T. Mitterling, B. Frauscher, S. V. Schönwald, and B. Högl, "Sleep modelled as a continuous and dynamic process predicts healthy ageing better than traditional sleep scoring," *Sleep Medicine*, vol. 77, pp. 136–146, jan 2021.

[145] L. W. Hermans, I. A. Huijben, H. van Gorp, T. R. Leufkens, P. Fonseca, S. Overeem, and M. M. van Gilst, "Representations of temporal sleep dynamics: Review and synthesis of the literature," *Sleep Medicine Reviews*, vol. 63, p. 101611, jun 2022.

[146] D. Shrivastava, S. Jung, M. Saadat, R. Sirohi, and K. Crewson, "How to interpret the results of a sleep study," *Journal of Community Hospital Internal Medicine Perspectives*, vol. 4, no. 5, p. 24983, jan 2014. [Online]. Available: /pmc/articles/PMC4246141//pmc/articles/PMC4246141/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4246141/

[147] A. S. Jordan, D. G. McSharry, and A. Malhotra, "Adult obstructive sleep apnoea," *The Lancet*, vol. 383, no. 9918, pp. 736–747, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0140673613607345

[148] K. B. Mikkelsen, H. Phan, M. L. Rank, M. C. Hemmsen, M. de Vos, and P. Kidmose, "Light-weight sleep monitoring: electrode distance matters more than placement for automatic scoring," Tech. Rep., 2017.

[149] M. Basunia, S. A. Fahmy, F. Schmidt, C. Agu, B. Bhattarai, V. Oke, D. Enriquez, and J. Quist, "Relationship of symptoms with sleep-stage abnormalities in obstructive sleep apnea-hypopnea syndrome," *Journal of Community Hospital Internal Medicine Perspectives*, vol. 6, no. 4, sep 2016. [Online]. Available: /pmc/articles/PMC5016742//pmc/articles/PMC5016742/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC5016742/

[150] M. T. Bianchi, S. S. Cash, J. Mietus, C. K. Peng, and R. Thomas, "Obstructive Sleep Apnea Alters Sleep Stage Transition

Dynamics," *PLoS ONE*, vol. 5, no. 6, 2010. [Online]. Available: /pmc/articles/PMC2893208//pmc/articles/PMC2893208/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC2893208/

[151] K. Shahveisi, A. Jalali, M. R. Moloudi, S. Moradi, A. Maroufi, and H. Khazaie, "Sleep Architecture in Patients With Primary Snoring and Obstructive Sleep Apnea," *Basic and Clinical Neuroscience*, vol. 9, no. 2, pp. 147–156, 2018. [Online]. Available: http://bcn.iums.ac.ir/article-1-895-en.html

[152] S. Redline, H. L. Kirchner, S. F. Quan, D. J. Gottlieb, V. Kapur, and A. Newman, "The Effects of Age, Sex, Ethnicity, and Sleep-Disordered Breathing on Sleep Architecture," *Archives of Internal Medicine*, vol. 164, no. 4, pp. 406–418, feb 2004. [Online]. Available: https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/216692

[153] M. Baumert, S. Hartmann, and H. Phan, "Automatic sleep staging for the young and the old – evaluating age bias in deep learning," *Sleep Medicine*, vol. 107, pp. 18–25, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1389945723001260

[154] A. Patanaik, J. L. Ong, J. J. Gooley, S. Ancoli-Israel, and M. W. L. Chee, "An end-to-end framework for real-time automatic sleep stage classification," *Sleep*, vol. 41, no. 5, p. zsy041, 03 2018. [Online]. Available: https://doi.org/10.1093/sleep/zsy041

[155] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning," *33rd International Conference on Machine Learning, ICML 2016*, vol. 3, pp. 1651–1660, jun 2015. [Online]. Available: https://arxiv.org/abs/1506.02142v6

[156] A. Stefani and B. Högl, "Sleep in Parkinson's disease," *Neuropsychopharmacology*, vol. 45, no. 1, p. 121, jan 2020. [Online]. Available: /pmc/articles/PMC6879568//pmc/articles/PMC6879568/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC6879568/

[157] D. L. Bliwise, M. L. Williams, D. Irbe, F. P. Ansari, and D. B. Rye, "Inter-rater Reliability for Identification of REM Sleep in Parkinson's Disease," *Sleep*, vol. 23, no. 5, pp. 1–6, 08 2000. [Online]. Available: https://doi.org/10.1093/sleep/23.5.1j

[158] H. Danker-Hopfe, D. Kunz, G. Gruber, G. Klösch, J. L. Lorenzo, S. L. Himanen, B. Kemp, T. Penzel, J. Röschke, H. Dorn, A. Schlögl, E. Trenker, and G. Dorffner, "Interrater reliability between scorers from eight european sleep laboratories in subjects with different sleep disorders," *Journal of*

Sleep Research, vol. 13, no. 1, pp. 63–69, 2004. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1365-2869.2003.00375.x

[159] R. Proost, L. Lagae, W. Van Paesschen, and K. Jansen, "Sleep in children with refractory epilepsy and epileptic encephalopathies: A systematic review of literature," *European Journal of Paediatric Neurology*, vol. 38, pp. 53–61, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1090379822000563

[160] A. A. Winsor, C. Richards, S. Bissell, S. Seri, A. Liew, and A. P. Bagshaw, "Sleep disruption in children and adolescents with epilepsy: A systematic review and meta-analysis," *Sleep Medicine Reviews*, vol. 57, p. 101416, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1087079221000010

[161] S. Y.-S. Chan, "Sleep architecture and homeostasis in children with epilepsy: a neurodevelopmental perspective," *Developmental Medicine & Child Neurology*, vol. 62, no. 4, pp. 426–433, 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/dmcn.14437

[162] Y. F. Lee, D. Gerashchenko, I. Timofeev, B. J. Bacskai, and K. V. Kastanenka, "Slow Wave Sleep Is a Promising Intervention Target for Alzheimer's Disease," *Frontiers in Neuroscience*, vol. 14, jun 2020. [Online]. Available: /pmc/articles/PMC7340158//pmc/articles/PMC7340158/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC7340158/

[163] A. Falter and M. J. A. V. D. Bossche, "How non-rapid eye movement sleep and Alzheimer pathology are linked," *World Journal of Psychiatry*, vol. 11, no. 11, p. 1027, nov 2021. [Online]. Available: /pmc/articles/PMC8613758//pmc/articles/PMC8613758/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC8613758/

[164] J. L. Ong, A. Patanaik, N. I. Y. N. Chee, X. K. Lee, J.-H. Poh, and M. W. L. Chee, "Auditory stimulation of sleep slow oscillations modulates subsequent memory encoding through altered hippocampal function," *Sleep*, vol. 41, no. 5, p. zsy031, 02 2018. [Online]. Available: https://doi.org/10.1093/sleep/zsy031

[165] H.-V. V. Ngo, T. Martinetz, J. Born, and M. Mölle, "Auditory closed-loop stimulation of the sleep slow oscillation enhances memory," *Neuron*, vol. 78, no. 3, pp. 545–553, 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0896627313002304

[166] M. M. Leminen, J. Virkkala, E. Saure, T. Paajanen, P. C. Zee, G. Santostasi, C. Hublin, K. Müller, T. Porkka-Heiskanen, M. Huotilainen, and T. Paunio, "Enhanced Memory Consolidation Via Automatic Sound Stimulation During Non-REM Sleep," *Sleep*, vol. 40, no. 3, mar 2017. [Online]. Available: /pmc/articles/PMC5806588//pmc/articles/PMC5806588/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC5806588/

[167] P. Scheltens, B. De Strooper, M. Kivipelto, H. Holstege, G. Chételat, C. E. Teunissen, J. Cummings, and W. M. van der Flier, "Alzheimer's disease," *The Lancet*, vol. 397, no. 10284, pp. 1577–1590, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0140673620322054

[168] M. Prince, A. Wimo, M. Guerchet, G.-C. Ali, Y.-T. Wu, and M. Prina, "World alzheimer report 2015. the global impact of dementia. an analysis of prevalence, incidence, cost and trends," 08 2015.

[169] B. P. Lucey, A. McCullough, E. C. Landsness, C. D. Toedebusch, J. S. McLeland, A. M. Zaza, A. M. Fagan, L. McCue, C. Xiong, J. C. Morris, T. L. S. Benzinger, and D. M. Holtzman, "Reduced non–rapid eye movement sleep is associated with tau pathology in early alzheimer's disease," *Science Translational Medicine*, vol. 11, no. 474, p. eaau6550, 2019. [Online]. Available: https://www.science.org/doi/abs/10.1126/scitranslmed.aau6550

[170] J. R. Winer, B. A. Mander, R. F. Helfrich, A. Maass, T. M. Harrison, S. L. Baker, R. T. Knight, W. J. Jagust, and M. P. Walker, "Sleep as a potential biomarker of tau and β-amyloid burden in the human brain," *Journal of Neuroscience*, vol. 39, no. 32, pp. 6315–6324, 2019. [Online]. Available: https://www.jneurosci.org/content/39/32/6315

[171] B. A. Mander, J. R. Winer, W. J. Jagust, and M. P. Walker, "Sleep: A novel mechanistic pathway, biomarker, and treatment target in the pathology of alzheimer's disease?" *Trends in Neurosciences*, vol. 39, no. 8, pp. 552–566, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0166223616300376

[172] A. S. Lim, M. Kowgier, L. Yu, A. S. Buchman, and D. A. Bennett, "Sleep Fragmentation and the Risk of Incident Alzheimer's Disease and Cognitive Decline in Older Persons," *Sleep*, vol. 36, no. 7, p. 1027, jul 2013. [Online]. Available: /pmc/articles/PMC3669060//pmc/articles/PMC3669060/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3669060/

[173] E. M. Ye, H. Sun, P. V. Krishnamurthy, N. Adra, W. Ganglberger, R. J. Thomas, A. D. Lam, and M. B. Westover, "Dementia detection from brain activity during sleep," *Sleep*, vol. 46, no. 3, p. zsac286, 11 2022. [Online]. Available: https://doi.org/10.1093/sleep/zsac286

[174] C. A. Derby, L. C. Burns, C. Wang, M. J. Katz, M. E. Zimmerman, G. L'italien, Z. Guo, R. M. Berman, and R. B. Lipton, "Screening for predementia AD: Time-dependent operating characteristics of episodic memory tests," *Neurology*, vol. 80, no. 14, p. 1307, apr 2013. [Online]. Available: /pmc/articles/PMC3656458//pmc/articles/PMC3656458/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3656458/

[175] S. Romanella, D. Roe, E. Tatti, D. Cappon, R. Paciorek, E. Testani, A. Rossi, S. Rossi, and E. Santarnecchi, "The sleep side of aging and alzheimer's disease," *Sleep Medicine*, vol. 77, pp. 209–225, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S138994572030232X

[176] A. D'Atri, S. Scarpelli, M. Gorgoni, I. Truglia, G. Lauri, S. Cordone, M. Ferrara, C. Marra, P. M. Rossini, and L. De Gennaro, "Eeg alterations during wake and sleep in mild cognitive impairment and alzheimer's disease," *iScience*, vol. 24, no. 4, p. 102386, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2589004221003540

[177] Y. Zhang, R. Ren, L. Yang, H. Zhang, Y. Shi, H. R. Okhravi, M. V. Vitiello, L. D. Sanford, and X. Tang, "Sleep in Alzheimer's disease: a systematic review and meta-analysis of polysomnographic findings," *Translational Psychiatry*, vol. 12, no. 1, dec 2022.

[178] P. Brayet, D. Petit, B. Frauscher, J.-F. Gagnon, N. Gosselin, K. Gagnon, I. Rouleau, and J. Montplaisir, "Quantitative eeg of rapid-eye-movement sleep: A marker of amnestic mild cognitive impairment," *Clinical EEG and Neuroscience*, vol. 47, no. 2, pp. 134–141, 2016, pMID: 26323578. [Online]. Available: https://doi.org/10.1177/1550059415603050

[179] A. K. F. Lam, J. Carrick, C.-H. Kao, C. L. Phillips, Y. Z. Zheng, B. J. Yee, J. W. Kim, R. R. Grunstein, S. L. Naismith, and A. L. D'Rozario, "EEG slowing during REM sleep in older adults with subjective cognitive impairment and mild cognitive impairment," *Sleep*, p. zsae051, 02 2024. [Online]. Available: https://doi.org/10.1093/sleep/zsae051

[180] Y. Ikoma, Y. Takahashi, D. Sasaki, and K. Matsui, "Properties of REM sleep alterations with epilepsy," *Brain*, vol. 146, no. 6, pp. 2431–2442, 03 2023. [Online]. Available: https://doi.org/10.1093/brain/awac499

[181] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[182] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, jun 2020.

[183] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52 138–52 160, sep 2018.

[184] F. Andreotti, H. Phan, and M. D. Vos, "Visualising convolutional neural network decisions in automated sleep scoring," in Proceedings of the *AIH@IJCAI*, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:52141352

[185] M. Dutt, S. Redhu, M. Goodwin, and C. W. Omlin, "SleepXAI: An explainable deep learning approach for multi-class sleep stage identification," *Applied Intelligence*, vol. 53, no. 13, pp. 16 830–16 843, jul 2022. [Online]. Available: https://link.springer.com/article/10.1007/s10489-022-04357-8

[186] H. Suresh, N. Hunt, A. Johnson, L. A. Celi, P. Szolovits, and M. Ghassemi, "Clinical intervention prediction and understanding with deep neural networks," in Proceedings of the *Machine Learning for Healthcare Conference*, pp. 322–337, 2017.

[187] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 4765–4774. [Online]. Available: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[188] S. Depeweg, J. M. Hernandez-Lobato, F. Doshi-Velez, and S. Udluft, "Decomposition of Uncertainty in Bayesian Deep Learning for Efficient and Risk-sensitive Learning," *35th International Conference on Machine Learning, ICML 2018*, vol. 3, pp. 1920–1934, oct 2017. [Online]. Available: https://arxiv.org/abs/1710.07283v4

[189] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," 12 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

[190] L. Deng, "The mnist database of handwritten digit images for machine learning research," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141–142, 2012.

Acknowledgments

This work was carried out at the ESAT-STADIUS Center for Dynamical Systems, Signal Processing, and Data Analytics (Department of Electrical Engineering), KU Leuven. Elisabeth Heremans was supported by the Research Foundation Flanders (FWO) (grant agreement No 1SC2921N).

Curriculum vitae



Elisabeth R.M. Heremans was born in Leuven, Belgium, on May 24, 1996. She earned her B.Sc. in Engineering with a major in Computer Science and her M.Sc. in Biomedical Engineering from KU Leuven, Belgium, in 2017 and 2019, respectively. Following a six-month research internship at École Polytechnique Fédérale de Lausanne, Switzerland, she began pursuing her Ph.D. in March 2020 at the STADIUS Center for Dynamical Systems, Signal Processing, and Data Analytics in KU Leuven's Electrical Engineering Department, under the supervision of prof. Maarten De Vos. During her Ph.D., Elisabeth completed a three-month research stay at the van der Schaar Lab at the University of Cambridge, as well as an internship at Microsoft Research. Her research interests include biomedical signal processing and machine learning, with a particular focus on brain signals and neuroscience applications.

List of Publications

Articles in internationally reviewed journals

1. **E. R. M. Heremans**, H. Phan, A. H. Ansari, P. Borzée, B. Buyse, D. Testelmans, and M. De Vos, "Feature matching as improved transfer learning technique for wearable EEG," *Biomedical Signal Processing and Control*, vol. 78, pp. 104009, 2022.

2. **E. R. M. Heremans**, H. Phan, P. Borzée, B. Buyse, D. Testelmans, and M. De Vos, "From unsupervised to semi-supervised adversarial domain adaptation in electroencephalography-based sleep staging," *Journal of Neural Engineering*, vol. 19, no. 4, pp. 036044, 2022.

3. L. Van den Bulcke, A. M. Peeters, **E. R. M. Heremans**, H. Davidoff, P. Borzée, M. De Vos, L. Emsell, J. Van den Stock, M. De Roo, J. Tournoy, B. Buyse, M. Vandenbulcke, C. Van Audenhove, D. Testelmans, and M. Van Den Bossche, "Acoustic stimulation as a promising technique to enhance slow-wave sleep in Alzheimer's disease: results of a pilot study," *Journal of Clinical Sleep Medicine*, vol. 19, no. 12, pp. 2107–2112, 2022.

4. H. Phan, K. P. Lorenzen, **E. R. M. Heremans**, O. Y. Chén, M. C. Tran, P. Koch, A. Mertins, M. Baumert, K. B. Mikkelsen, and M. De Vos, "L-SeqSleepNet: whole-cycle long sequence modeling for automatic sleep staging," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 10, pp. 4748-4757, 2023.

5. **E. R. M. Heremans**, N. Seedat, B. Buyse, D. Testelmans, M. van der Schaar, and M. De Vos, "U-PASS: an Uncertainty-guided deep learning Pipeline for Automated Sleep Staging," *Computers in Biology and Medicine*, vol. 171, pp. 108205, 2024.

6. **E. R. M. Heremans**, L. Van den Bulcke, N. Seedat, A. Devulder, P. Borzée, B. Buyse, D. Testelmans, M. Van Den Bossche, M. van der Schaar, and M. De Vos, "Automated remote sleep monitoring needs uncertainty quantification," *Journal of Sleep Research*, e14300, 2024.

7. R. Proost, **E. R. M. Heremans**, L. Lagae, W. Van Paesschen, M. De Vos, and K. Jansen, "Automated sleep staging on reduced channels in children with epilepsy," *Frontiers in Neurology*, vol. 15, pp. 1390465, 2024.

8. K. P. Lorenzen, **E. R. M. Heremans**, M. De Vos, and K. B. Mikkelsen, "Personalization of automatic sleep scoring: How best to adapt models to personal domains in wearable EEG," Accepted in *IEEE Journal of Biomedical and Health Informatics*, 2024.

9. L. Van den Bulcke, H. Davidoff, **E. R. M. Heremans**, Y. Potts, K. Vansteelandt, M. De Vos, D. Christiaens, L. Emsell, L. H. Jacobson, D. Hoyer, B. Buyse, M. Vandenbulcke, D. Testelmans, and M. Van Den Bossche, "Acoustic stimulation to improve slow-wave sleep in Alzheimer's disease: a multiple night at-home intervention," *The American Journal of Geriatric Psychiatry*, 2024.

Articles submitted for review

1. **E. R. M. Heremans***, A. Devulder*, P. Borzée, R. Vandenberghe, F.-L. De Winter, M. Vandenbulcke, M. Van Den Bossche, B. Buyse, D. Testelmans, W. Van Paesschen, and M. De Vos, "Home-based wearable sleep recording augmented by artificial intelligence for Alzheimer's disease screening: a method-development and proof-of-concept study," Submitted for review, 2024. (*: joint first authors)

2. J. Macea Ortiz, **E. R. M. Heremans**, R. Proost, M. De Vos, and W. Van Paesschen, "Automated sleep staging using deep learning on wearable EEG and accelerometry in people living with epilepsy," Submitted for review, 2024.

3. A. Devulder, G. Vanderlinden, **E. R. M. Heremans**, J. Schaeverbeke, L. Van Langenhoven, D. Testelmans, A. Kalkanis, M. Van Den Bossche, F.-L. De Winter, M. Vandenbulcke, R. Vandenberghe, K. Van Laere, M. De Vos, W. Van Paesschen, "Associations between sleep and early regional tau pathology," Submitted for review, 2024.

Papers in proceedings of international conferences

1. **E. R. M. Heremans**, T. Osselaer, N. Seeuws, H. Phan, D. Testelmans, and M. De Vos, "Data augmentation in semi-supervised adversarial domain adaptation for EEG-based sleep staging," in Proceedings of the *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, Ioannina, Greece, pp. 1-4, 2022.

2. H. Phan, **E. R. M. Heremans**, O. Y. Chén, P. Koch, A. Mertins and M. De Vos, "Improving automatic sleep staging via temporal smoothness regularization," in Proceedings of the *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, pp. 1-5, 2023.

3. **E. R. M. Heremans** and M. De Vos, "Explaining uncertainty in AI for clinical decision support systems," Accepted in Proceedings of the *Uncertainty meets Explainability Workshop at ECML-PKDD 2023*, Torino, Italy, pp. 1-9, 2023.

Abstracts in proceedings of (inter)national conferences

1. **E. R. M. Heremans** and M. De Vos, "Domain adaptation for sleep staging on wearable EEG data," *19th National Day of Biomedical Engineering*, Online Conference, Belgium, 27 November 2020. Oral presentation.

2. **E. R. M. Heremans**, H. Phan, A. H. Ansari, and M. De Vos "Feature matching: a novel domain adaptation technique for sleep staging," *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, Virtual Conference, 27-30 July 2021. Oral presentation.

3. **E. R. M. Heremans**, "Automated sleep analysis toward sleep monitoring at home," *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, 2-10 June 2023. Poster presentation.

4. **E. R. M. Heremans**, A. Devulder, B. Buyse, D. Testelmans, W. Van Paesschen, and M. De Vos, "Automatic sleep staging and detection of sleep disorders through wearable EEG monitoring devices," *Sleep Medicine*, 115, S231-S232, 2024. (Abstract presented at *World Sleep Congress 2023*, Rio De Janeiro, Brazil, 20-25 October 2023.) Oral presentation.

5. L. Micca, **E. R. M. Heremans**, D. Testelmans, P. Borzée, W. Vandenberghe, M. De Vos, and M. Gilat, "AI-driven automated sleep staging in Parkinson's disease", *27th Congress of the European Sleep Research Society*, Seville, Spain, 2024.

Workshop at international conference

1. M. De Vos, **E. R. M. Heremans**, N. Seeuws, H. Phan and A. Gramfort, "EEG Workshop – EEG signal processing for the future: integrating insights across domains," Workshop organized at *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, 5 June 2023. Workshop proposal, workshop organization, webpage creation, and oral presentation.

Software

1. **E. R. M. Heremans** and M. De Vos, "Feature matching for sleep staging networks," 2023. [Online]. Available at `https://github.com/elisabethRMH/featurematching`
2. **E. R. M. Heremans** and M. De Vos, "Unsupervised and semi-supervised adversarial domain adaptation," 2023. [Online]. Available at `https://github.com/elisabethRMH/adversarial_DA`
3. **E. R. M. Heremans** and M. De Vos, "U-PASS: An uncertainty-guided deep learning pipeline for automated sleep staging," 2023. [Online]. Available at `https://github.com/elisabethRMH/U-PASS`

Science communication and outreach

1. **E. R. M. Heremans** "De AI die jou zal helpen slapen," *Science Figured Out*, September 2023. Video pitch.



FACULTY OF ENGINEERING SCIENCE
DEPARTMENT OF ELECTRICAL ENGINEERING
STADIUS CENTER FOR DYNAMICAL SYSTEMS, SIGNAL PROCESSING AND DATA ANALYTICS
Kasteelpark Arenberg 10 box 2446
B-3001 Leuven
elisabeth.heremans@kuleuven.be
http://esat.kuleuven.be/stadius