

Artificial Intelligence in Regulated Domains: Lessons from Sleep Medicine

Emil Harðarson

Dissertation submitted to the School of Computer Science
at Reykjavik University in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

June, 2026

Supervisors:	María Óskarsdóttir, Associate Professor, University of Southampton and Reykjavík University
	Anna Sigríður Islind, Professor, Reykjavik University
	Erna Sif Arnardóttir, Associate Professor, Reykjavik University
Examining Board:	Henri Korkalainen, Senior Researcher, University of Eastern Finland
	Hafsteinn Einarsson, Associate Professor, University of Iceland
Examiner:	Federico Cabitza, Associate Professor, University of Milan-Bicocca

© Emil Harðarson
June, 2026

Artificial Intelligence in Regulated Domains: Lessons from Sleep Medicine

Emil Harðarson

PhD Thesis
Department of Computer Science
Reykjavik University

ISBN (print version): 9978-9935-577-06-1
ISBN (electronic version): 9978-9935-577-07-8
Author's ORCID: 0000-0002-9497-4265

This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0).

© 2026 Emil Harðarson

Artificial Intelligence in Regulated Domains: Lessons from Sleep Medicine

Emil Harðarson

April, 2026

Abstract

This thesis investigates how artificial intelligence (AI) and machine learning (ML) can be integrated into the regulated workflow of polysomnography-based sleep architecture analysis while ensuring outcomes are understandable and defensible. Rather than focusing solely on sleep-stage classification accuracy, the thesis argues that different parts of the workflow require different combinations of machine learning, explicit rules, uncertainty representation, and human oversight.

Across six papers, the thesis develops AI methods for sleep analysis and, through simulation and empirical studies, examines how their outputs behave within the clinical workflow. These include an LLM-guided neural architecture search framework for autonomously developing multimodal sleep staging models; a rule-based automatic sleep staging system derived from clinical scoring rules, capable of producing auditable justifications in clinical language; simulation and empirical studies of how sleep staging error propagates to downstream sleep metrics (such as total sleep time and time spent in deep sleep); and methods for estimating sleep metrics without performing sleep stage classification.

The results show that strong epoch-level agreement does not translate uniformly into reliable downstream sleep metrics and that some sleep metrics vary substantially across expert scorers. They further show that preserving probabilistic uncertainty can improve sleep metric estimation, that direct sleep metric estimation can be competitive with conventional hypnogram-based pipelines, and that parts of the formal scoring workflow can be translated into executable and auditable procedures. Interviews with sleep professionals also show that effective AI decision support requires interactivity, familiar clinical language, and system designs that preserve the human expert's accountability.

These findings are synthesized in the MARBLE framework

(Multiple Abstraction-Level Rule-Based Learning Engine), which treats AI decision support in sleep medicine as the coordinated use of rule-based and pattern-based methods across multiple abstraction levels, from raw physiological signals and micro-events to sleep stages, sleep metrics, and higher-level interpretation. The thesis concludes that AI integration in clinical sleep analysis should not be framed as the automation of one prediction task, but as the careful design of decision support across a rule-governed workflow. Although developed in sleep medicine, this may also have implications for AI decision support in other regulated domains.
Keywords: Regulated AI, explainable AI, decision support systems, sleep staging, polysomnography, neural architecture search.

Acknowledgements

More people contributed to this work than I can name here. This thesis has truly been a team effort, and I am grateful to everyone involved.

My deepest gratitude goes to my supervisor, María Óskarsdóttir, whose influence on this work, and on me as a researcher, cannot be overstated. She brought scientific rigor and a sharp editorial hand to every draft, and she was always available when needed. She gave me the freedom to follow research ideas that often led nowhere but sometimes yielded the most interesting findings. Whatever depth this thesis has, it grew from that patience and flexibility. I could not have asked for a steadier hand to guide my work.

I would like to express my sincere gratitude to my co-supervisors, Anna Sigríður Islind and Erna Sif Arnardóttir, for their support. I am grateful for the time they dedicated to regular meetings and for their thorough feedback on my papers and drafts. Anna Sigga's encouragement to engage more deeply with information systems broadened the scope of my research in ways I did not anticipate. Erna's deep expertise in sleep research gave my work the scientific grounding it needed. Balancing the demands of the Sleep Revolution and other projects, they maintained their commitment to my work, and I am fortunate to have had their guidance.

I thank Hafsteinn Einarsson and Henri Korkalainen for their careful and thorough review of the thesis. Their excellent suggestions greatly improved this work, and I am grateful for the diligence they brought to serving on my PhD committee. I also thank my examiner, Federico Cabitza, for his careful reading of the thesis and the thoughtfulness of his examination. I am also grateful for his advice earlier in the process, which helped sharpen my thinking about ground truth and inter-rater agreement.

My thanks to the co-authors of the papers: Gabriel Jouan, Konstantin Popov, Luka Biedebach, Frida Ivarsson, Sigríður Sigurðardóttir, Ómar Bessi Ómarsson, and Teitur Hrólfsson, for their collaboration and the many great conversations.

I owe a great deal to Gabriel, whose generosity as a colleague made the PhD journey richer. Working alongside him sharpened my understanding of deep learning and probability theory, among much else, and his willingness to collaborate, explain, and explore ideas together left a big mark.

Particular thanks also go to Konstantin for our countless conversations, which have given me a perspective I could not have found elsewhere.

To all the members of the Sleep Revolution team, Kristín Anna Ólafsdóttir, Heiður Grétarsdóttir, Katrín Hera Gústafsdóttir, Danielle Knieriem, Conor Murphy, and the many others, thank you for your support and companionship throughout this project. This includes my fellow PhD students, Eysteinn Finnsson, Benedikt Hólm Þórðarson, Elena Richert, Bjarki Freyr Sveinbjarnarson, Dimitri Ferretti, Lisa Schmitz, Jacopo Piccini, Kolfinna Þórisdóttir, and all the others.

I am grateful for the support of the Department of Computer Science, in particular Luca Aceto and Jacky Mallett, for their encouragement and guidance. My thanks also to Verity Louise Sharp and everyone else in the office for helping me navigate the logistics of the PhD program.

Thank you to the students whose projects contributed to the research: Marie Raffray, Juliette Allia, Sigurður Baldursson, and Bent Gunnarsson.

I was fortunate to visit the STAR team at the University of Eastern Finland on two occasions, and I am grateful to Timo Leppänen, Samu Kainulainen, Matias Rusanen, Sami Nikkonen, and everyone else there for their warmth, generosity, and collaboration.

Thank you to Jón Skírnir Ágústsson and the team at Nox Medical for the valuable conversations and assistance.

I would also like to thank Kristján Rúnar Kristjánsson and Ingvar Ingvarsson at Íslandsbanki, and Sindri Reynisson at Landsbankinn, for their time and for helping me think through the potential applications of MARBLE in mortgage decision support and risk assessment.

To my family and friends, far too numerous to name in full, thank you for your support in all its forms. My thanks, in par-

ticular, to my parents, Hörður and Fanney; my brothers, Þorri and Fannar; and my in-laws, Gunnar, Camilla, Emil Gunnar, and Edda Laufey.

To my children, Jakob, Lísbet, and Matthías, thank you for keeping me grounded throughout. And to Edda, thank you for everything. This is as much your work as it is mine.

Finally, a special thanks to my grandparents, Þór Sigurbjörnsson and Þuríður Björnsdóttir, who supported my PhD studies both directly and indirectly. I know I will never be able to repay their kindness and generosity, but I will do my best to pass it on to future generations. *This thesis is dedicated to them.*

This thesis was supported by the European Union's Horizon 2020 research and innovation program under grant agreement 965417 (the Sleep Revolution project) and the Icelandic Research Fund under Doctoral Student Grant 2410607-051, and the Department of Computer Science at Reykjavík University. Mobility grants from the Erasmus+ program facilitated research visits, and the Icelandic Student Innovation Fund supported a related student project. I am grateful for the support of the taxpayers who make these public funds possible.

The creation of the thesis and the research it contains were aided by generative AI. This includes the use of systems such as OpenAI GPT models, Google Gemini, Semantic Scholar, and Grammarly. AI tools were used for tasks such as brainstorming ideas, finding research articles, phrasing of text, reviewing text, writing and debugging code, and more. In fact, I struggle to think of a single substantial component of this work that was not in some way supported by an AI system. In keeping with the findings and spirit of the thesis, AI was used responsibly and with strict human oversight.

Publications list

Included publications

Publication I

(Manuscript in preparation.) E. Hardarson *et al.*, 2026, *Data-Local Autonomous LLM-Guided Neural Architecture Search for Multiclass Multimodal Time-Series Classification*, DOI: 10.48550/arXiv.2603.15939

Publication II

(Submitted to Journal of Sleep Research.) E. Hardarson *et al.*, 2026, *Staging by the Book: Automatic Sleep Stage Classification Using Scoring Rules*, DOI: 10.48550/ARXIV.2605.22859

Publication III

E. Hardarson *et al.*, 2023, "Error Propagation from Sleep Stage Classification to Derived Sleep Parameters in Machine Learning on Data from Wearables", *Current Sleep Medicine Reports*, vol. 9, DOI: 10.1007/s40675-023-00253-w

---

Publication IV

(Manuscript in preparation.) E. Hardarson *et al.*, 2026, "Inter-scorer reliability in sleep architecture metrics"

Publication V

(Manuscript in preparation.) E. Hardarson *et al.*, 2026, "From sleep staging to sleep metrics: Rethinking targets in automated sleep analysis"

Publication VI

E. Hardarson *et al.*, 2024, "Human-AI Collaboration: From Explainable AI to Co-Creating Meaning", *ACIS 2024 Proceedings*

Other publications

B. Holm *et al.*, 2024, "An optimized framework for processing multicentric polysomnographic data incorporating expert human oversight", *Frontiers in Neuroinformatics*, vol. 18, DOI: 10.3389/fninf.2024.1379932

List of Tables

1.1 Mapping from thesis contributions to research questions 6

3.1 Declaration of Authorship Contribution 27

4.1 Benchmark overview for the LLM-guided NAS . . 40

5.1 Development-record confusion matrix in percentages. 65
5.2 Development-record confusion matrix (epoch counts). 66
5.3 Test-set confusion matrix in percentages. 66
5.4 Test set confusion matrix (epoch counts). 67
5.5 Method vs. human mean absolute differences from consensus sleep metrics. 68

6.1 Summary of published ML sleep staging studies using wrist-worn wearables. 78
6.2 Example generated confusion matrix with 80% accuracy and low REM sensitivity. 81
6.3 Example generated confusion matrix with 80% accuracy and low light-sleep sensitivity 82
6.4 Mean and standard deviation of cells in 20,000 generated confusion matrices. 82
6.5 Reported and simulated RMSE for sleep metric error residuals. 87
6.6 Performance of baseline models for staging and sleep metric estimation. 92

7.1 Cohort characteristics of the SRMS dataset. 98
7.2 Inter-scorer agreement coefficients for epoch-by-epoch sleep staging. 107

7.3 Inter-scorer agreement coefficients for derived sleep metrics. 108
7.4 Leave-one-scorer-out error and comparison with a naive constant predictor. 110
7.5 Variance decomposition of derived sleep metrics across recordings and scorers. 116
7.6 Comparison of two consensus methods for sleep metrics. 117

8.1 Dataset characteristics for SHHS and Sleep Revolution cohorts. 129
8.2 Approximate parameter counts for the evaluated models. 140
8.3 Sleep metric estimation results on the SHHS dataset. 145

9.1 Examples of explanations shown in the XAI interface. 165
9.2 Qualities of XAI where the target audience is domain experts 169

A.1 Pooled Bland–Altman summary statistics from scorer-versus-leave-one-out comparisons. 246
A.2 Summary of pairwise absolute disagreement for derived sleep metrics across all scorer pairs and recordings. 248

List of Figures

1.1 Hierarchical framing of the thesis problem, application domain, thesis questions, and research questions. 5

2.1 Example hypnogram 10
2.2 Number of publications about automatic sleep staging. 13
2.3 Accuracy of automatic sleep staging methods. . . . 13

3.1 Design Science Research Framework. 24

4.1 Overview of the expert decomposition and representation-level fusion. 35
4.2 Schematic of the data-local autonomous NAS loop. 37
4.3 Search trajectories for binary experts on the SleepEDFx dataset NAS. 43
4.4 Search trajectories for binary experts on the Ethanol-Concentration dataset NAS. 43

5.1 Overview of the rule-based sleep staging process. . 53
5.2 Example epoch with micro-annotations across EEG, EOG, and EMG signals. 60
5.3 Example of a sequential elimination trace produced by the rule-based sleep staging method. 61
5.4 Example of a natural-language explanation dialogue produced by the rule-based sleep staging method . 62
5.5 Development-record hypnogram: consensus vs. predicted, with disagreements highlighted. 63
5.6 Test-record hypnogram: consensus vs. predicted, with disagreements highlighted. 64
5.7 Human agreement distribution for epochs where the algorithm agrees vs. disagrees with consensus. . . . 69

6.1 Schematic of the simulation process for estimating sleep metric error. 83
6.2 Time-of-night distribution of sleep stages in the training set. 85
6.3 Effect of resampling a hypnogram using published confusion matrices. 86
6.4 Simulated distributions of sleep metric error across the dataset. 88
6.5 Sleep metric RMSE as a function of staging error. 91

7.1 Hypnograms and derived sleep metrics for one subject across 10 scorers. 113
7.2 Distribution of scorer-derived sleep metrics across recordings. 114
7.3 Relationship between staging disagreement and error in sleep metrics. 115

8.1 Clinical sleep analysis workflow and pathways for bypassing the hypnogram. 126
8.2 Hard vs. soft aggregation of hypnodensity for sleep metric estimation. 135
8.3 Overview of the heuristic signal intensity (HSI) model. 136
8.4 Vision Transformer (ViT) architecture used for direct estimation of sleep metrics. 141
8.5 Sleep metric estimation error results on the test set of the SHHS dataset. 149
8.6 Example HSI model output with hypnogram and probability traces. 150
8.7 Leave-one-out comparison of sleep metric errors against human consensus. 151

9.1 Transparent rule-based model as an inherently explainable system. 159
9.2 Black-box ML with post-hoc explanation as strategic framing. 159
9.3 Co-creating meaning through human-AI collaboration within a rule-based framework. 162
9.4 A screenshot of the interview interface. 163

10.1 MARBLE overview for sleep analysis 178
10.2 SleepScorer report-level overview. 180
10.3 SleepScorer parameter-level view. 181

10.4 SleepScorer hypnogram-level view. 182
10.5 SleepScorer epoch-level view. 183
10.6 Chat interface for discussing decisions in the SleepScorer system. 184

11.1 A rule-based decision-making procedure at multiple abstraction levels. 193
11.2 Multi-level mortgage decision pipeline for a financial institution in Iceland 194
11.3 Mortgage application overview in a demonstration interface for MARBLE. 195
11.4 Credit score view in the MARBLE mortgage DSS . 195

A.1 Bland–Altman plots for selected derived sleep metrics using scorer-versus-consensus comparisons . . . 245

Abbreviations

AASM American Academy of Sleep Medicine.

AI Artificial Intelligence.

CDSS Clinical Decision Support Systems.

DL Deep Learning.

DSR Design Science Research.

DSRM Design Science Research Methodology.

DSS Decision Support System.

EDF European Data Format.

EEG Electroencephalogram.

EHR Electronic Health Record.

EMG Electromyogram.

EOG Electrooculogram.

HSI Heuristic Signal Intensity.

LAMF Low Amplitude Mixed Frequency.

LLM Large Language Model.

LTV loan-to-value.

MARBLE Multiple Abstraction-Level Rule-Based Learning Engine.

ML Machine Learning.

N1 Non-REM sleep stage 1.

N2 Non-REM sleep stage 2.

N3 Non-REM sleep stage 3.

NA Number of Awakenings.

NAS Neural Architecture Search.

NREM Non-Rapid Eye Movement.

PSG Polysomnography.

R&K Rechtschaffen and Kales.

REM Rapid Eye Movement.

REMs Rapid Eye Movements.

SE Sleep Efficiency.

SHHS Sleep Heart Health Study.

SOL Sleep Onset Latency.

SRMS Sleep Revolution Multicentre Scoring Dataset.

TRT Total Recording Time.

TST Total Sleep Time.

UI User Interface.

ViT Vision Transformer.

WASO Wake After Sleep Onset.

XAI Explainable Artificial Intelligence.

Contents

Abstract iii
Acknowledgements v
Publications list ix
List of Tables xi
List of Figures xii

1 Introduction **1**
1.1 Thesis aims and contribution 2
1.2 A few words on vocabulary 7

2 Background and related work **9**
2.1 Sleep medicine as the application domain 9
2.1.1 Sleep architecture 9
2.1.2 Automatic sleep staging 11
2.2 Conceptual background for regulated AI 14
2.2.1 Expert systems, clinical decision support systems and symbolic AI 15
2.2.2 Machine learning and neural AI 16
2.2.3 Interpretability, explainability and justifiability 17
2.2.4 Abstraction hierarchies and the relevance problem 18
2.2.5 Hybrid and neurosymbolic AI 19

3 Research approach **23**
3.1 Datasets 25
3.2 Researcher's role 26

4 LLM-guided neural architecture search **29**
4.1 Introduction 29
4.2 Related Work 32

4.3 Method . . . 33
4.3.1 Fusion network of modular binary experts . 33
4.3.2 LLM-guided neural architecture search for expert models . . . 35
4.3.3 Experimental setup and benchmarks . . . 37
4.4 Results . . . 38
4.4.1 Experimental overview . . . 38
4.4.2 Baseline model . . . 42
4.4.3 Effect of LLM-guided NAS . . . 42
4.4.4 Overall takeaways . . . 44
4.5 Discussion . . . 44
4.5.1 Open Access Tool . . . 44
4.5.2 Data Privacy . . . 45
4.5.3 Limitations . . . 45
4.6 Conclusion . . . 46

5 Staging by the book **49**
5.1 Introduction . . . 49
5.2 Method . . . 52
5.2.1 Algorithm . . . 52
5.2.2 Explanation generation . . . 55
5.2.3 Dataset . . . 57
5.2.4 Evaluation . . . 58
5.3 Results . . . 59
5.3.1 Output data . . . 59
5.3.2 Comparison to human consensus . . . 60
5.3.3 Assessment of discrepancies . . . 61
5.4 Discussion . . . 70
5.5 Conclusion . . . 71

6 Error propagation to sleep metrics **73**
6.1 Introduction . . . 73
6.2 Background . . . 76
6.2.1 Sleep metrics . . . 77
6.3 Methods . . . 79
6.3.1 Generation of Confusion Matrices . . . 80
6.3.2 Baseline models . . . 83
6.4 Results and discussion . . . 85
6.4.1 Baseline values . . . 86
6.4.2 Clinical relevance of staging error . . . 89
6.5 Conclusions . . . 93

7 Inter-scorer reliability in sleep architecture metrics 95
7.1 Introduction 95
7.2 Methods 97
7.2.1 Dataset and scoring protocol 97
7.2.2 Definitions of sleep architecture metrics . . 98
7.2.3 Staging agreement 100
7.2.4 Agreement of sleep architecture metrics . . 101
7.2.5 Leave-one-scorer-out error 102
7.2.6 Naive constant predictor comparison 103
7.2.7 Variance decomposition 104
7.2.8 Consensus construction comparison 104
7.2.9 Bootstrap confidence intervals 105
7.2.10 Software and implementation 106
7.3 Results 106
7.3.1 Epoch-by-epoch staging agreement 106
7.3.2 Agreement of derived sleep metrics 107
7.3.3 Leave-one-scorer-out error 109
7.3.4 Variance decomposition 111
7.3.5 Consensus construction comparison 111
7.4 Discussion 118
7.5 Conclusion 122

8 Rethinking prediction targets in automated sleep analysis 125
8.1 Introduction 125
8.2 Methods 128
8.2.1 Datasets 128
8.2.2 Hypnogram-based sleep metrics 129
8.2.3 Estimation families 131
8.2.4 Sleep staging models 132
8.2.5 Soft aggregation of hypnodensity 134
8.2.6 Direct estimation models 135
8.2.7 Experiments 142
8.2.8 Evaluation and statistical analysis 143
8.3 Results 144
8.3.1 Preserving uncertainty improves sleep metric estimation 144
8.3.2 Direct estimation of sleep metrics is competitive 146
8.3.3 Cross-dataset performance is limited 146

8.3.4 Direct sleep metric estimation under scorer disagreement . . . 147
8.4 Discussion . . . 152
8.5 Conclusion . . . 155

9 Human-AI collaboration 157
9.1 Introduction . . . 157
9.2 Explainability in AI literature . . . 159
9.3 A framing perspective on XAI . . . 160
9.4 Research method . . . 162
9.4.1 Research context . . . 162
9.4.2 Data collection and analysis . . . 163
9.5 Explainability in clinical practice . . . 165
9.5.1 Rule-based models or pattern recognition . 166
9.5.2 Interactivity and multi-way communication 166
9.5.3 Communication via familiar language . . . 167
9.5.4 Transparency of model internals . . . 167
9.6 Discussion . . . 168
9.7 Conclusion . . . 171

10 The MARBLE framework 173
10.1 MARBLE . . . 173
10.1.1 Abstraction levels and pathways . . . 174
10.1.2 Flags as the primary output . . . 175
10.1.3 Why parallel pathways at every level . . . 175
10.1.4 Relation to prior hybrid and neurosymbolic systems . . . 176
10.2 SleepScorer: MARBLE applied to sleep analysis . . 179
10.2.1 System architecture . . . 179
10.2.2 Navigating the abstraction hierarchy . . . 179
10.2.3 The chat assistant . . . 184
10.2.4 Known limitations . . . 185

11 Discussion 187
11.1 Key findings . . . 187
11.2 Answering the research questions . . . 190
11.3 Limitations . . . 191
11.4 Looking beyond sleep . . . 192
11.5 Future work . . . 196

12 Conclusion 199

Bibliography **201**

A Paper IV supplement **243**
A.1 Bland–Altman analysis 243
A.2 Pairwise sleep-metric disagreement 247
A.2.1 Pairwise scorer-versus-scorer sleep-metric disagreement 247

Chapter 1

Introduction

Artificial intelligence (AI) is being deployed across professional practice faster than institutions can develop policies, norms, and procedures for its use [1], [2]. The gap is widest in regulated domains, such as medicine, finance, and law, where decisions are not only judged by whether they are correct, but also by whether they can be reviewed, audited, and justified according to accepted rules. In such settings, an AI system may be technically successful in the narrow sense of producing accurate predictions while being practically useless, because its outputs are not connected to the procedures by which decisions are normally made. The questions of where AI should enter a workflow, what kinds of outputs it should produce, and how its conclusions can remain accountable to the rules already in place, are therefore pressing [3].

This thesis studies these questions in the setting of clinical sleep architecture analysis. Sleep disorders affect cardiovascular health, cognition, mental health, and daytime safety, and are highly prevalent in the general population [4]–[6]. Polysomnography (PSG), the overnight recording of physiological signals during sleep, is the principal technique used to diagnose sleep disorders and investigate sleep architecture [7]. PSG interpretation is rule-governed: trained experts apply standardized scoring rules [8] to derive sleep stages, sleep metrics, and a clinical report.

Sleep architecture analysis is small enough in scope to study end-to-end, yet institutionally complex enough to expose the integration challenges that arise when AI is introduced into a regulated clinical workflow. The dominant framing in the AI sleep

architecture analysis literature treats the problem as supervised classification of sleep stages. That is, given physiological signals, the task is to predict the sleep stage labels a human expert would assign, and measure progress by agreement with those labels [9]. This framing, while useful for benchmarking, is misaligned with how sleep analysis functions in clinical practice. Sleep stage labels are intermediate outputs. They are inputs to sleep metrics and diagnosis. It is the downstream artifacts, not the sleep stage labels themselves, that carry clinical weight.

1.1 Thesis aims and contribution

The primary objective of this thesis is to investigate how AI and ML can be integrated into the complex, regulated workflow of sleep architecture analysis without sacrificing the justifiability of outcomes. We operationalize justifiability as follows: an outcome is justifiable in terms of existing rules if an independent domain expert can (1) trace the outcome to explicit rule applications and the underlying evidence used to satisfy the rules, and (2) reproduce the decision path from the recorded inputs.

The research examines technical and clinical limitations of current automatic sleep staging methods and proposes alternatives. The work has four aims:

> **Aim 1:** Characterize the clinical and technical limitations of conventional ML sleep staging.
>
> **Aim 2:** Identify alternative prediction targets for ML sleep architecture analysis.
>
> **Aim 3:** Determine to what extent clinical sleep scoring rules can be formalized and where ML can support them.
>
> **Aim 4:** Define the requirements for human-AI collaboration in clinical sleep architecture analysis.

This thesis examines the problem of integrating AI into the workflow of sleep architecture analysis through specific thesis questions: in PSG sleep architecture analysis, what decisions should be automated with ML, and how should those decisions be justified? Although the work presented is spread across six papers that address distinct technical topics, they are unified by

those questions. The papers discuss topics in sleep staging, such as how far statistical pattern recognition can be pushed, how much of the scoring rules can be operationalized into computer code, and whether conventional prediction targets and explanation paradigms are appropriate for use in clinical practice. Together, they build toward a view of ML automation in which predictive performance, clinical relevance, and justifiability are solved jointly in a decision support system. The four aims of the thesis can be formulated as the following research questions:

> **RQ 1:** Under what conditions do conventional ML sleep staging models and evaluation metrics (accuracy, Cohen's kappa) fail to be clinically useful?
>
> **RQ 2:** Are there targets and evaluation metrics that are more robust and clinically relevant than epoch-by-epoch sleep stage classification if the clinical task is interpreting sleep architecture?
>
> **RQ 3:** To what extent can clinical sleep scoring rules be translated into an executable procedure, and where do ML models provide support within that procedure?
>
> **RQ 4:** What must a clinical sleep architecture analysis decision support system provide for human experts to collaborate with AI models?

Across six papers and a prototype decision support system, the thesis contributes both technical artifacts and design knowledge about how to use ML models inside a rule-governed clinical workflow without sacrificing justifiability. The relationship between the overarching thesis problem, the application domain of sleep architecture analysis, the thesis questions, and the four research questions is shown in Figure 1.1. The contributions of the thesis and their relation to the papers and research questions are also summarized across the six papers in Table 1.1.

The remainder of the thesis is structured as follows:

- Chapter 2 provides background on sleep architecture and automatic sleep staging, as well as AI in regulated domains.
- Chapter 3 describes the research approach, datasets, and the role of the researcher.

- Chapter 4 presents Paper I, which develops an LLM-guided neural architecture search framework for multimodal sleep staging.
- Chapter 5 presents Paper II, which translates parts of the AASM sleep scoring procedure into an executable rule-based sleep staging method.
- Chapter 6 presents Paper III, which studies how sleep staging errors propagate into downstream sleep metrics.
- Chapter 7 presents Paper IV, which examines inter-scorer reliability in derived sleep architecture metrics.
- Chapter 8 presents Paper V, which evaluates alternatives to hypnogram-based sleep metric estimation.
- Chapter 9 presents Paper VI, which studies human-AI collaboration and explainability in clinical sleep analysis.
- Chapter 10 synthesizes the papers into the MARBLE framework and presents the SleepScorer prototype.
- Chapters 11 and 12 discuss the overall findings, implications, future work, and conclusions.

The thesis has no single, unified methods chapter. Instead, the methods are described within each paper chapter (Chapters 4–9). This is due to the methodological diversity of the work: the papers use distinct approaches, including neural architecture search, rule-based system design, simulation studies, statistical reliability analysis, model evaluation, and qualitative interview analysis.

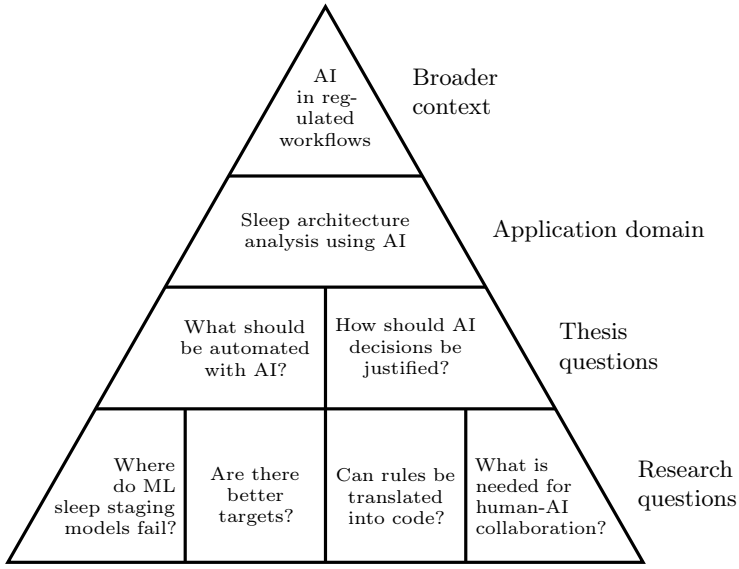


Figure 1.1: Hierarchical framing of the thesis problem, application domain, thesis questions, and research questions.

Table 1.1: Mapping from thesis contributions to research questions (RQ1–RQ4) and their role in the thesis-level goal of supporting integration of machine learning into a rule-governed clinical workflow.

	Contribution	**RQs**	**Link to thesis**
Ch. 4	A multimodal sleep staging model discovered via a novel NAS framework.	RQ1	Establishes a high-performing ML sleep staging model based purely on statistical pattern recognition.
Ch. 5	An automatic sleep staging algorithm based on AASM scoring rules that provides justifications using clinical terminology.	RQ3	Demonstrates the feasibility of translating parts of the clinical scoring procedure into executable, auditable code.
Ch. 6	A simulation framework for propagating staging error into errors in derived sleep metrics.	RQ1, RQ2	Shows that conventional staging metrics can fail to reflect downstream clinical relevance.
Ch. 7	An empirical analysis of inter-scorer reliability in derived sleep architecture metrics.	RQ1, RQ2	Shows that sleep metric targets are themselves uncertain and that the construction of the reference value matters.
Ch. 8.3	Methods for estimating sleep metrics without relying on the hypnogram, including soft aggregation and direct estimation.	RQ2, RQ3	Provides alternatives to epoch-by-epoch staging that better match the clinical task of interpreting sleep architecture and help reposition where ML is most useful in the workflow.
Ch. 9	A study of requirements for sleep architecture analysis decision support and a framing of human-AI collaboration.	RQ4	Specifies what experts need to remain accountable humans-in-the-loop: interactive, rule-aligned decision support.
Ch. 10	The MARBLE framework, which synthesizes the thesis findings into a multi-level framework for rule-governed AI decision support in sleep analysis.	RQ1–RQ4	Integrates the findings of the six papers and presents a framework for AI decision support across multiple abstraction levels.

1.2 A few words on vocabulary

Several terms used throughout this thesis are unavoidably polysemous. In particular, there are some words that carry established, but very different meanings in sleep medicine and ML literature. This is unfortunate because the overlap is frequent and occurs in both clinical and technical parts of the text. There are a few especially problematic and commonly occurring terms:

> **Epoch.** In sleep medicine, an *epoch* refers almost exclusively to a 30-second time segment of a PSG recording. In ML, an *epoch* usually refers to a single full pass of a neural network through a training set. Throughout this thesis, *epoch* refers to a 30-second segment of a sleep recording, whereas a neural network pass through a training set is referred to as a *training epoch*.
>
> **Parameters.** In ML, *model parameters* are numerical values learned during training. In sleep medicine, *sleep parameters* are quantitative descriptors derived from the PSG, such as total sleep time. To avoid confusion, we refer to these as *sleep architecture metrics*, or simply *sleep metrics*.
>
> **Architecture.** In sleep medicine, *sleep architecture* refers to the structure of sleep across the night. In ML, *network architecture* describes the structure, parameters, and hyperparameters of an ML model. Throughout this thesis, the term *architecture* is used with an explicit prefix, i.e. *sleep-* or *network architecture*, unless the intended meaning is clear from the context.

This thesis is largely about finding a *neural* network *architecture* with different model *parameters* and hyperparameters (e.g., the number of training *epochs*), to predict *epochs* and other sleep *architecture parameters* using *neural* signals. Such is life.

Chapter 2

Background and related work

This chapter establishes the background of the thesis in two sections. It first introduces sleep medicine as the application domain, and in particular, the hierarchical structure of sleep architecture analysis and the current state of automatic sleep staging. It then turns to the broader literature on AI in regulated domains, including the engineering tradition of encoding clinical rules as executable code, machine learning, explainable AI, the relevance problem in hierarchical decision systems, and hybrid approaches that combine rule-based and pattern-based reasoning.

2.1 Sleep medicine as the application domain

2.1.1 Sleep architecture

In clinical sleep medicine and sleep research, sleep stage scoring serves as an important diagnostic tool. It is the process by which the different stages of sleep are identified and categorized, most commonly following clinical guidelines established by the American Academy of Sleep Medicine (AASM) [8]. These stages are wake, rapid eye movement sleep (REM), and non-rapid eye movement sleep stages 1, 2, and 3 (N1, N2, and N3). The classification of the sleep stages is based on physiological signals captured during sleep,

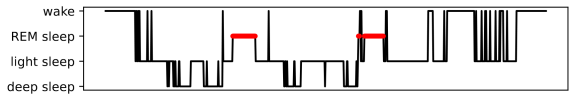


Figure 2.1: An example of a hypnogram. REM sleep is traditionally shown in red.

such as electroencephalography (EEG), electrooculography (EOG), and chin electromyography (EMG). Whereas EEG measures the brain's electrical activity, EOG records eye movements, and chin EMG measures muscle activity in the chin [7], [8].

These measurements are typically conducted during a PSG sleep study, a comprehensive test used to measure sleep and diagnose sleep disorders. During a PSG, these physiological signals and others are recorded to provide a detailed overview of an individual's sleep patterns.

For the purpose of analysis, a night of sleep is divided into 30-second segments, known as epochs.[1] Each epoch is then classified into one of the five sleep stages. This classification, which is called sleep stage scoring, is performed manually by trained sleep technologists during a process that typically takes between one and three hours [11], [12].

From the results of the epoch-by-epoch sleep stage scoring, a range of sleep metrics can be calculated. These include, for example, the total sleep time (TST), which measures the total duration of sleep; sleep efficiency (SE), the percentage of time in bed spent sleeping; and the sleep onset latency (SOL), the time it takes to fall asleep. Additionally, the analysis yields a hypnogram, a graphical representation of the transitions between sleep stages over the course of the night. Figure 2.1 shows an example of a hypnogram. This visual summary provides insight into the architecture and quality of sleep [8], [13].

Rechtschaffen and Kales (R&K) published the first standardized sleep scoring manual in 1968 [14]. The scoring rules have since been updated, and the AASM scoring manual now serves as the standard for scoring sleep studies, including how to assign sleep stages [8]. According to the AASM scoring manual, for example, stage N1 is identified by the attenuation of wake-related alpha

[1]The reason for this choice of time duration of epochs is largely historical: 30 seconds was approximately the length of the printer paper during early analog PSG recordings [10].

rhythm and its replacement by low-amplitude mixed-frequency (LAMF) EEG activity, that is, relatively low-voltage EEG without a single dominant rhythm. In individuals who do not clearly generate alpha rhythm, N1 may instead be recognized by features such as vertex sharp waves (V-waves; brief, sharply contoured waveforms maximal over central EEG derivations) or slow eye movements (rolling eye movements seen in the EOG) [8]. Learning to score sleep stages, therefore, requires the scorer not only to apply formal rules but also to recognize recurring physiological patterns such as V-waves, slow eye movements, and LAMF activity. Scoring also requires interpretation of the rules since the rules are ambiguous and fail to cover all scenarios that occur in practice [15]. Commonly, colleagues teach each other ways of interpreting the rules, resulting in "house rules" [15], [16]. Expert human sleep scorers typically agree on sleep stage classification for around 86% of epochs (inter-scorer agreement) [17]. Intra-rater agreement (how often a scorer agrees with their own scoring) is around 88-90% [18].

The analysis of sleep architecture can be viewed as a hierarchy of processes and decisions. At the lowest level, events such as spindles and V-waves are recognized. One level up the hierarchy, these *micro-annotations* determine which sleep stage each 30-second epoch should be assigned. The epoch-by-epoch sleep stage assignments form the hypnogram, from which sleep metrics are calculated. At the highest level of the hierarchy, sleep metrics, together with the hypnogram and other clinical information, are assembled into a sleep report, which is used for the diagnosis of the patient.

2.1.2 Automatic sleep staging

Automated sleep staging methods have been proposed for decades but have largely failed to gain widespread adoption in clinical practice [15], [19]. Early approaches attempted to encode the human scoring procedure as algorithms [20], [21]. These systems depended on hand-crafted features along with rule-based decision trees or thresholding to reproduce the procedures of the R&K scoring manual [14]. Throughout the 1990s and early 2000s, a range of expert systems for sleep staging were proposed [21].

One prominent example was the Somnolyzer 24×7 system, which was first created to emulate the R&K rules and later modified

to comply with the AASM rules [22]–[24]. Another example of a rule-based automatic sleep staging method was Liang *et al.* [25], which used 14 "if-then" rules inspired by the AASM scoring manual.

By the 2010s, the focus of research in automated sleep staging had shifted to supervised ML algorithms [26], [27]. Early ML methods typically involved feature engineering and training classifiers such as support vector machines or random forests [28]. With the rise of DL in the late 2010s, architectures such as convolutional neural networks and recurrent neural networks have been used to classify sleep stages with minimally preprocessed signals. Models such as DeepSleepNet [29], SeqSleepNet [30], and U-Sleep [31] have shown sleep staging performance exceeding that of feature-engineered systems and even human inter-scorer agreement.

There have also been some hybrid approaches to automatic sleep staging. In the year 2000, Park *et al.* [32] proposed a hybrid neural network and rule-based expert system for automatic sleep stage classification. More recently, in 2022, Hwang *et al.* [33] presented an AI-based clinical decision support system for assisting sleep technologists in reviewing AI-predicted sleep stages, designing the system around clinically relevant features such as sleep spindles, K-complexes, and delta waves.

Figure 2.2 shows the number of publications about automatic sleep staging (blue line) and AI sleep staging (red dashed line) published per year up to the year 2024. In total, there are 525 articles about automatic sleep staging, 288 (55%) of which were published in the last 5 years. The data comes from a PubMed search of

```
automatic AND (("sleep stage") OR
("sleep staging"))
```

and

```
(("machine learning") OR
("artificial intelligence")) AND
(("sleep stage") OR ("sleep staging"))
```

respectively.

Figure 2.3 shows the accuracy of the automatic sleep staging methods reported in the papers. The data was collected by retrieving the abstracts of the papers [22], [34]–[107] from the PubMed search and collecting the reported 5-class sleep staging classification accuracy. The figure also shows three horizontal lines,

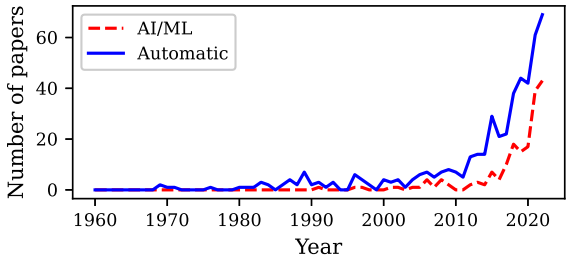


Figure 2.2: The number of publications about automatic sleep staging (blue line) and AI sleep staging (red dashed line) published per year.

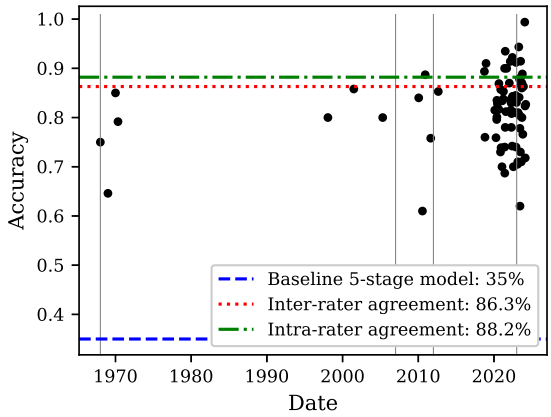


Figure 2.3: The accuracy of the automatic sleep staging methods reported in the papers, over time. Gray vertical lines show years where changes were made in the scoring rules: R&K and AASM scoring manual versions 1, 2, and 3. The blue dashed line shows the baseline accuracy of an educated random guess, the red dotted line shows inter-rater agreement, and the green line shows intra-rater agreement [17], [18]

representing the baseline accuracy of an educated random guess (see Chapter 6), inter-rater agreement [17], and intra-rater agreement [18], from top to bottom, respectively. Any automatic sleep staging method that lies above the top line has likely been overfitted since it replicates the sleep stage scoring of a human scorer more closely than human scorers reproduce their own scoring, on average. The figure shows that the accuracy of the automatic sleep staging methods has been close to the inter-rater agreement for the past five decades.

Despite strong reported performance in the automatic sleep staging literature, clinical adoption has remained limited. There are disagreements in the literature about the current state of automatic sleep staging. Some authors claim that advancements in ML and the availability of large datasets have improved automatic sleep staging in the last few years [108], to the point of having "basically been solved" [9]. Others argue that little substantial progress has been made. According to Stanley [15], we have "imperfect staging rules being imperfectly implemented by dozens of companies in closed source algorithms". To navigate this uncertainty, researchers and clinicians are increasingly calling for human-in-the-loop approaches to AI sleep staging [109]–[111]. Human-in-the-loop AI refers to systems where human expertise is integrated into the AI algorithms during training or inference. This discussion mirrors the broader calls for human-in-the-loop AI in healthcare [112].

2.2 Conceptual background for regulated AI

The automated analysis of sleep architecture is a specific instance of a general task: the deployment of AI in regulated decision processes. The questions this thesis asks about sleep architecture analysis are not unique to sleep medicine, and the literature that has engaged the questions most directly comes from outside of the field of sleep. This section explores some of the relevant literature, starting with expert systems, clinical decision support systems, and symbolic AI.

2.2.1 Expert systems, clinical decision support systems and symbolic AI

In the 1960s and 1970s, early expert systems supported decision-making by operationalizing domain expertise as explicit rules. Dendral [113] and MYCIN [114] are often cited early examples of expert systems for scientific hypothesis formation and medical decision support, respectively. A defining feature of these systems is their ability to produce reasoning traces of the form "rule X fired because evidence Y was present". Expert systems are a type of *symbolic AI* [115], [116], which are systems that represent knowledge explicitly as rules or logic and perform inference over those representations.

A closely related field in medicine is that of clinical decision support systems (CDSS). Clinical decision support systems are software systems designed to assist clinicians in making decisions about diagnosis, treatment, or monitoring. Many CDSS are implemented as rule-based systems, encoding clinical knowledge in the form of explicit code. The longest-running standard for representing executable clinical knowledge is the Arden Syntax [117], [118]. Knowledge is encoded in Medical Logic Modules (MLMs), each of which combines a *data* slot specifying which observations the module depends on, a *logic* slot containing the rule itself, and an *action* slot specifying what to do when the rule fires. MLMs were designed to be human-readable and authored by domain experts rather than programmers, with the goal that a rule written at one institution could be shared and executed at another. Efforts have been undertaken to standardize the representation of data in CDSS, such as through the OMOP Common Data Model maintained by the OHDSI collaborative [119].

The modern HL7 stack pulls encoding, data, and invocation apart and standardizes each separately. Fast Healthcare Interoperability Resources (FHIR) [120], [121] provides a REST/JSON resource data model; Clinical Quality Language (CQL) [122] expresses clinical logic against a FHIR data model; SMART on FHIR [121] specifies how substitutable clinical applications launch inside an EHR; and CDS Hooks [123] specifies how an external decision-support service is invoked from within the clinician's workflow. Recent surveys [124], [125] document growing activity around these FHIR-based CDS standards since 2018. Sleep medicine does not have a comparable data-standards layer. How-

ever, the AASM rules can be implemented as executable logic, as shown in Chapter 5.

Purely symbolic approaches run into a number of problems when applied to real-world data. Firstly, they are *brittle*. Rules do not cover all the infinite number of borderline cases and exceptions that can appear in an open-ended environment [126]. Secondly, symbolic rules assume clean inputs, while data from the real world are noisy and contain missing values [127]. Furthermore, real-world systems are in constant evolution, and symbolic AI requires ongoing maintenance and validation. This is referred to as the "knowledge acquisition bottleneck" [128]. These problem areas for symbolic AI are where ML, or *neural AI*, excels.

2.2.2 Machine learning and neural AI

Machine learning approaches the task from the opposite starting point. Rather than representing domain knowledge as explicit rules, ML systems learn statistical patterns from data. In supervised learning, a model is trained on examples consisting of inputs and target outputs, and the learned model is then used to predict outputs for new inputs [129]. Supervised ML includes a range of model families. Some methods, such as logistic regression, support vector machines, random forests, and gradient-boosted trees, are often used with hand-crafted features extracted from the input data. Deep learning methods, such as convolutional, recurrent, and transformer-based neural networks, instead learn internal representations from the data itself.

However, these data-driven ML models introduce a problem of their own. Their predictions are often difficult to trace to accepted clinical rules, and their internal representations are not usually inspectable in the way that a rule base is [130], [131]. For example, an ML model may predict a sleep stage with high accuracy without providing a decision path that a sleep technologist can reproduce from the underlying signal evidence. The same statistical flexibility that makes ML useful for noisy pattern recognition also makes it harder to audit, justify, and govern. This is referred to as the "black-box" problem in AI [132]

2.2.3 Interpretability, explainability and justifiability

Explainable AI (XAI) is a field of research that aims to make black-box ML systems more transparent and understandable to humans. However, there is not a single agreed-upon definition of what an "explanation" is in the XAI literature [133]. This lack of agreement is not unique to AI: the concept of *explanation* itself has long been the subject of philosophical discussion [134], [135]. In practice, XAI is often framed as a way to increase trust in AI systems by providing explanations [136].

XAI is commonly divided into two strategies [130], [137]. The first, interpretability, treats transparency as a property of the model itself: simple linear models, decision trees, and rule sets expose their reasoning by construction. The second, post-hoc explainability, accepts model opacity and reconstructs an account of the prediction after the fact, typically through attribution methods applied to a trained black-box model. Chapter 9 returns to this distinction in the context of human–AI collaboration; for the present chapter it is enough that the post-hoc strand has been the subject of sustained criticism. Rudin [138] argues that post-hoc explanations are themselves unreliable and potentially misleading, and advocates for interpretable models rather than retrofitted explanations. Babic and Cohen [139] go further, calling post-hoc explanations "insincere-by-design" on the grounds that they cannot reveal the reasoning that actually produced a model's output. Together, these critiques motivate a shift from asking whether a model can be explained to whether an AI-enabled decision can be justified. Following Babic and Cohen [139], *justifiability* concerns whether a decision can be defended to relevant stakeholders by appeal to accepted grounds (e.g., rules or procedures), rather than by a persuasive narrative about what the model might have attended to. Babic and Cohen [139] argue that post-hoc explanations are a poor route to justifiability. Because they are generated after the fact, post-hoc XAI methods can return arbitrary rationales, and they do not reliably distinguish permissible justifications from spurious or illegitimate ones.

As AI systems grow increasingly powerful in their predictive capability, their outputs become more difficult to explain. If we require faithful, mechanistic explanations, then producing such explanations effectively requires access to (and a model of) the AI's

decision procedure. For complex systems, that kind of explanation may be as complicated as the system itself [131]. Explanations must thus bring relevant facts to the foreground, while suppressing countless other, less relevant facts.

2.2.4 Abstraction hierarchies and the relevance problem

Section 2.1.1 introduced sleep architecture analysis as a hierarchical workflow. Physiological signals are interpreted as micro-events, micro-events inform epoch-by-epoch sleep stage scoring decisions, epochs form a hypnogram, the hypnogram is used to derive sleep metrics, and these metrics contribute to the sleep report. This workflow is a hierarchy, and supporting decisions with AI requires choosing where in the hierarchy questions should be answered. The same is true of explanations. To explain a decision is not to list every fact that could have influenced it, but to select the level at which the decision should be made understandable [140].

This connects the thesis to the *relevance problem*. The relevance problem concerns how an intelligent agent (human or artificial) manages to restrict its thinking and perception to what matters, at the "correct" level of abstraction. Each decision such an agent makes is affected by an effectively unbounded number of facts. If an agent tried to consider every fact that could affect its decision, it would never finish reasoning, but if it fails to consider an important fact, the decision will be irrational.

To make the relevance problem concrete, we can imagine an AI agent tasked with assisting a park ranger at the entrance to a public park.[2] This seems like a simple task. When faced with the question "is this allowed in the park?", the AI agent needs to (1) understand what *this* is and (2) what rules apply to *this*. Even before any explicit reasoning begins, the AI agent needs to decide what information to retrieve, what features to attend to, and which *frame* to use, whether the relevant concept is *vehicle*, *maintenance equipment*, *weapon*, *emergency*, *alcohol*, and so on.

The rule "no vehicles in the park" [141] is clear for automobiles and motorcycles, but less so for bicycles, wheelchairs, airplanes flying overhead, and so on. On one level, ambiguity can be eliminated by adding definitions (*vehicles* are a means of carrying or

[2]This example is inspired by H. L. A. Hart's famous "no vehicles in the park" example, more on that later.

transporting something [142]), but each definition carries with it its own set of borderline cases, and new circumstances bring novel examples that were not anticipated. Complete determinancy would require an endless expansion of subrules and exceptions [141].

The theory of *relevance realization* [143] provides a useful vocabulary for this movement between levels. According to this theory, relevance is not a fixed property attached to individual facts, but an ongoing process of framing and reframing a situation. When a coarse representation does not support a satisfactory decision, an agent may shift to a finer-grained representation; when local details are no longer useful, the agent may move back to a higher-level view. In the context of sleep architecture analysis, this means moving between report-level interpretation, sleep metrics, the hypnogram, disputed epochs, and the underlying physiological signal evidence.

This framing motivates the multi-level approach developed later in the thesis. A clinically useful AI system for sleep architecture analysis should not be understood as a single predictor operating at one fixed level. It should instead support decision making across levels of abstraction, combining pattern-based methods for noisy evidence with rule-based structures for auditable decision making. This provides the conceptual basis for the MARBLE framework developed in Chapter 10.

2.2.5 Hybrid and neurosymbolic AI

The preceding subsections point to a tension: encoding clinical rules as executable code makes decisions auditable but brittle; data-driven ML handles the brittleness but introduces the black-box problem; and decisions are made across a hierarchy of abstraction levels at which both rules and pattern recognition have roles to play. Hybrid AI, or *neurosymbolic AI* is an engineering response to this [144]. The goal is to build hybrid AI systems where ML is used for pattern recognition and symbolic reasoning is used for structured decision processes.

In clinical decision systems, Kierner *et al.* [145] proposes a taxonomy of neurosymbolic systems, identifying five variants: rules embedded in ML (REML), ML pre-processes input for rule-based inference (MLRB), rule-based pre-processing for ML prediction (RBML), rules influence ML training (RMLT), and parallel ensem-

bles of rules and ML (PERML).

Kautz [146] proposes a different taxonomy of six types of neurosymbolic system architecture. One variant, which Kautz calls "Symbolic Neuro symbolic" or "symbols in/out with neural in the middle," describes systems where a learned model performs pattern recognition, extracting structured facts from noisy inputs, and a symbolic layer applies decision logic over those facts. In Kierner et al.'s terminology, this corresponds to MLRB. Prenosil *et al.* [147] is an example of this type of neurosymbolic AI. They connect GPT-4 to a rule-based expert system, using the LLM to extract facts from natural-language reports and the expert system to verify medical rules. Their system produces deterministic labels and an auditable reasoning trace.

Many variants of hybrid AI architectures do not follow this clean pipeline structure of ML, producing inputs that are handed off to a rule engine. Some methods have symbolic logic calling multiple neural components as subroutines, or they shape the training and inference through rules and constraints. Kauffman *et al.* [148] illustrate this by performing probabilistic inference on EHR-derived temporal graphs while using deep geometric learning to obtain the representations used in the inference. Ma *et al.* [149] incorporate medical rules as posterior regularization to push learned risk predictions towards behavior consistent with known rules. In Kierner *et al.* [145]'s taxonomy, this falls under RMLT. La Cava *et al.* [150] use symbolic regression to construct a clinical prediction of hypertension classification from EHR records. Symbolic regression is a supervised ML technique that finds mathematical (symbolic) models that fit the data, rather than learning model parameters from data [151].

Although not explicitly described in Kautz's taxonomy, one possible implementation of neurosymbolic AI is one of *redundancy*, where neural and symbolic pathways are run in parallel. In Kierner *et al.* [145]'s taxonomy, this is classified as PERML. Examples of this include Corny *et al.* [152], where an ML component prioritizes high-risk prescriptions while a rule-based check and human oversight provide a safety layer. Kierner *et al.* [153] provides their own PERML fusion example in vaccine-hesitancy prediction.

In sleep analysis, the Park *et al.* [32] hybrid discussed in Section 2.1.2 is an early PERML-style system: neural networks are invoked when the rule-based system fails to reach a deterministic decision, with the two pathways effectively voting on disputed epochs. Of

the patterns in the taxonomy, PERML is the closest category to the architecture developed in this thesis: MARBLE (Chapter 10) runs rule-based and pattern-based pathways in parallel, but applies the redundancy across multiple abstraction levels rather than operating at a single prediction point.

In summary, sleep architecture analysis (Section 2.1.1) is a rule-governed clinical workflow organized as a hierarchy of abstraction levels, and automatic sleep staging research has oscillated between rule-based and pattern-based approaches. Encoding clinical scoring rules is necessary but not sufficient (Section 2.2.1); ML handles the brittleness of symbolic logic but introduces the black-box problem (Section 2.2.2); relevance realization across the decision hierarchy is a proposed cognitive structure for navigating decisions (Section 2.2.4); and hybrid neurosymbolic system architectures run rule-based and pattern-based processing in parallel (Section 2.2.5). The remainder of this thesis develops this position into a decision support system for sleep architecture analysis.

Chapter 3

Research approach

The research in this thesis broadly follows the Design Science Research (DSR) paradigm. DSR is a research methodology in which knowledge is produced by designing and studying artifacts that solve identified problems [154], [155]. Rather than treating systems, models, methods, and prototypes as engineering artifacts, DSR treats them as research contributions. The main requirement is that an artifact addresses a clearly formulated problem and that claims about its utility are supported by empirical evidence.

Design Science Research Methodology (DSRM) is a commonly used process model for conducting DSR. In DSRM, research is structured into six activities [155], [156]:

1. *Problem identification and motivation*: define the problem and why it matters.
2. *Definition of objectives for a solution*: specify "success".
3. *Design and development*: build an artifact that targets an objective.
4. *Demonstration*: show that the artifact can be applied to the problem.
5. *Evaluation*: measure performance against objectives.
6. *Communication*: report the problem, artifact, and evidence to stakeholders.

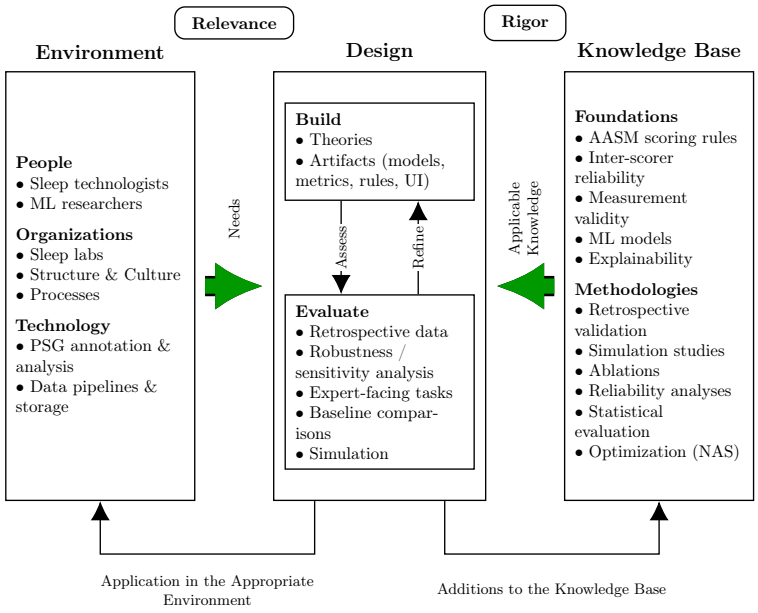


Figure 3.1: Design Science Research Framework. The figure is from Hevner *et al.* [154], adapted to fit the focus of the thesis.

This thesis adopts an iterative version of DSRM [155] and is grounded in the exploratory analysis of failure modes. The focus is on recurring mismatches between how ML sleep staging methods are developed and evaluated on the one hand and how they are used clinically on the other. The work began from the assumption that automatic sleep staging from PSG signals was technically solved, but through collaboration with clinical sleep experts, it became clear that the clinical value of sleep staging lies primarily in understanding sleep architecture and derived sleep metrics rather than epoch labels. This directed the thesis work towards understanding how ML sleep staging can hold utility within the broader clinical workflow.

The artifacts developed in this thesis include: ML models for sleep staging and sleep metric estimation; new evaluation metrics and validation procedures; simulations of error propagation from sleep staging to sleep metrics; executable scoring rules; and UI prototypes for sleep analysis decision support systems.

3.1 Datasets

The research in this thesis uses three datasets: the Sleep Heart Health Study (SHHS) dataset [157], [158], the SleepEDFx dataset [159], and the Sleep Revolution Mulicentre Scorer (SRMS) dataset [17]. Apart from their availability, these datasets were chosen because they serve three distinct purposes: large-scale training data (SHHS); accessible open data (SleepEDFx); and human consensus analysis (SRMS).

SHHS is a large, multi-center prospective cohort study in which participants underwent overnight PSG. The SHHS cohort includes data from over 6000 participants. In this thesis, SHHS was used for neural architecture search of sleep staging models and direct estimation of sleep metrics, because of the large number of participants.

SleepEDFx is an openly available dataset of 197 whole-night PSG recordings, widely used for benchmarking in the sleep staging literature. It contains EEG, EOG, chin EMG, and event markers manually scored according to the R&K manual. In this thesis, the 197 human-scored hypnograms were used in the simulations of Paper III.

The SRMS dataset consists of 50 PSG recordings scored in-

dependently by 10 human expert scorers, enabling analysis of inter-scorer agreement in sleep metrics in Paper IV. It was also used to evaluate models in Papers II and V.

3.2 Researcher's role

The research in this thesis was enabled by collaborating with sleep technologists and other clinicians. I had hands-on experience collecting PSG data, observed sleep scoring sessions, and had numerous in-depth discussions about the nuances of sleep scoring with experienced scorers. This position made it possible to observe recurring themes, including a gap between what automatic sleep staging methods optimize for and what clinicians need. This is the origin of the thesis's central claim: that automatic sleep staging should be treated as one component in a regulated workflow where justifications play a central role and are not treated as an afterthought.

This embedded perspective introduces the risk of biasing the problem formulation and proposed solutions to a specific clinical environment. Local workflow conventions and "house-rules" can bias the problem formulation and proposed solutions towards a specific clinical environment [15], [16]. This risk was partly mitigated by working within the Sleep Revolution project, which allowed me to collaborate with experts from sleep centers across multiple countries. The Sleep Revolution is a multicenter international initiative that brought together sleep experts and data from several sleep centers across different countries [160].

Table 3.1 summarizes my contribution to each paper using the ME/EE/CE labels as per Reykjavik University rules.

Table 3.1: Declaration of Authorship Contribution (ME = Main Effort, EE = Equal Effort, CE = Contributing Effort)

	Idea	Related work & literature	Data gathering	Research design	Artifact design	Analysis & synthesis	Draft	Admin-istration
Paper 1	ME	EE	-	ME	ME	ME	ME	ME
Paper 2	ME	ME	CE	ME	ME	ME	ME	ME
Paper 3	ME	ME	-	ME	ME	ME	ME	ME
Paper 4	ME	ME	CE	ME	EE	ME	ME	ME
Paper 5	ME	EE	CE	ME	ME	ME	ME	ME
Paper 6	ME	ME	ME	ME	ME	ME	ME	ME

The 'Data gathering' column is '–' for Papers I–III because those studies rely on existing PSG datasets and previously collected annotations rather than collecting new empirical material within this thesis.
In Papers II, IV, and V, I contributed to data gathering (CE) through work related to the multi-scorer SRMS material and its use in the study.
In Paper VI, I led data gathering (ME) because the work is based on interviews and iterative feedback with prospective clinical users.

Chapter 4

LLM-guided neural architecture search

Data-local autonomous LLM-guided neural architecture search for multiclass multimodal time-series classification

Full reference: (Manuscript in preparation.) E. Hardarson *et al.*, 2026, *Data-Local Autonomous LLM-Guided Neural Architecture Search for Multiclass Multimodal Time-Series Classification*, DOI: 10.48550/arXiv.2603.15939

4.1 Introduction

Motivation. Designing preprocessing pipelines and model architectures for classifying multiclass, multimodal time-series data is a challenge for machine learning engineers. Multimodal time series data refers to multiple synchronized streams of temporal data of different modalities [161]. These temporal modalities may include, for example, sensor measurements, audio, video, or physiological data [162]. With the increasing presence of smartphones, digital-

ized manufacturing, wearable devices, smart home sensors, and medical devices, an abundance of multimodal data is frequently produced and is becoming increasingly complex [163]. Potential uses of this data include healthcare [164], activity recognition [165], and the Internet of Things [166]. Physiological sleep monitoring is a prime example of multimodal time-series classification, as it involves synchronized signals from different modalities [167], which is why we use it as an application domain for our proposed system.

With increasing numbers of modalities, experimentation becomes more complex. They pose a particularly challenging design space because (i) modalities are heterogeneous and provide complementary evidence, (ii) class structure induces non-uniform confusion patterns, and (iii) architectural choices interact with modality handling and fusion strategy [161]. Consequently, the standard workflow of manually iterating on architectures and training protocols can be both slow and brittle. AutoML tools, which automate feature engineering, model selection, hyperparameter tuning, and evaluation, promise a way to overcome this challenge [168]. Particularly, Neural Architecture Search (NAS), introduced in 2016 by Zoph and Le, specifically searches the architecture space of artificial neural networks [169].

Problem Statement. Nowadays, NAS and related automated machine learning techniques have become standard tools for reducing the manual effort of model design. However, in many applied settings, the dominant bottleneck is not the absence of optimization primitives, but the practical burden of *running* long-horizon experimentation: configuring and restarting runs, diagnosing failures, tracking intermediate artifacts, and deciding what to try next under limited compute resources. Furthermore, many domains that produce multimodal data, such as healthcare, are prohibited from applying these tools, since sensitive, personal data cannot be moved to an external tool or server for optimization and training. In such settings, progress is often limited by *human-hours per unit of experimental progress*, even when substantial compute time is acceptable.

Proposed Solution. To address these challenges, we propose an autonomous, data-local experimentation framework that formalizes NAS-driven experimentation as a *research cycle*: the system repeatedly proposes candidate expert architectures, instantiates

training runs, evaluates outcomes, and records results as persistent artifacts. The framework is designed to reduce the number of human hours spent on orchestration and iteration while allowing long-running, compute-intensive search processes to run unattended. To make the multiclass multimodal structure explicit, we adopt a modular decomposition in which we train one binary expert per class and per modality, and then learn a lightweight fusion model over the experts' penultimate representations.

Another novelty of our proposed solution is a large-language-model (LLM)-guided optimization approach. We employ an LLM as a controller to guide the neural architecture search for the individual expert models, proposing candidate architectures based on observed validation performance and training dynamics from prior evaluations. All model training and evaluation are performed locally under a fixed protocol, with the LLM receiving only aggregated metrics and artifacts, enabling iterative architecture optimization while preserving data locality and privacy.

Contributions. We present and evaluate the proposed framework and provide a proof-of-concept in two settings. First, by comparing a simple baseline NAS with existing classifiers in UEA30, a benchmarking dataset of multimodal classification tasks, and SleepEDFx, a sleep classification benchmarking data set. Second, by showing the enhanced performance of the LLM-guided NAS. Existing solutions have shown to be effective for one-dimensional data, but multimodal time series data has, to the best of our knowledge, not been approached with AutoML tools yet. In summary, our main contributions are the following:

- We introduce an autonomous, data-local experimentation framework for NAS that reduces human-hours by formalizing experimentation as a research cycle with persistent, auditable artifacts.
- We operationalize multiclass multimodal learning via a modular expert decomposition: one binary classifier per class and per modality, with a trainable fusion MLP over penultimate representations; NAS is applied only to the experts.
- We provide empirical evidence on 22 multiclass multivariate time-series tasks and demonstrate the effectiveness of our approach on two of them.

The rest of this chapter is organized as follows. In the next section, we discuss research that is related to our work. In Section 4.3, we present the two parts of our proposed methodology: the fusion network of binary experts and the LLM-guided NAS. Section 4.4 presents the results of our experiments, and in Section 4.5 we discuss their implications, e.g. regarding data privacy. Finally, we summarize our work and contributions in Section 4.6.

4.2 Related Work

Various AutoML [170] and NAS [171] tools have been proposed in existing literature. Our contribution to this existing body of work is by combining several research threads: multimodal time-series classification and fusion, multiclass decomposition and expert-style modeling, NAS for temporal and multimodal architectures, practical AutoML and long-horizon experimentation, and the emerging use of LLMs as assistants in machine learning workflows. The unifying motivation is to support systematic model exploration for multiclass multimodal time series under realistic operational constraints, with an emphasis on reducing human effort in iterative experimentation.

AutoML for Ensemble Learning. A central design dimension in multimodal learning is the fusion strategy: early fusion combines modalities at the input or low-level feature stage, intermediate fusion combines learned representations (often through concatenation, attention, or gating), and late fusion combines decisions (e.g., logits or probabilities) produced by modality-specific models [172]. We propose an architecture that creates separate models for each modality, which is in itself not a novel approach, as it has been applied, for example, by Bastos et al. [173], but we add to existing literature by developing a system to automatically optimize these separate models. Erickson et al. introduced AutoGluon, an AutoML tool that automatically optimizes the building of multiple ensemble models [174]. McKinnon and Atkinson developed a search space for multivariate time-series NAS [175], but no existing tool addresses the specific needs of multimodal time-series data.

LLM-guided AutoML. A recent thread in AutoML explores using LLMs to assist or guide NAS, motivated by the observation

that many architecture changes are naturally expressed, for example, as structured edits to code or high-level design constraints. The natural language used in these models is a fundamental difference to traditional AutoML tools, as Miguel-Morante et al. showed with their AutoML tool, emphasizing the human-centric approach of providing ML systems in a more intuitive and accessible way and strengthening the role of human-computer interaction [176]. Existing LLM-to-NAS methods vary in how they control representation and feedback. Trirat, Jeong, and Hwang, for example, created an agent-based LLM-guided AutoML tool that takes users' task descriptions and facilitates collaboration among specialized LLM agents to deliver deployment-ready models [177].

Data-Local AutoML. Our work adds to existing research and existing AutoML software, and, furthermore, from a data-privacy perspective. Since our architecture allows data to be stored locally, improving the model iteratively solely by receiving textual summary information about the data and reviewing training statistics ensures full data privacy and protection. Other AutoML tools, such as H2O [178] and AutoSklearn [179], support a fully local implementation, which also ensures privacy. Our solution, however, preserves this local property while still leveraging the capabilities of remotely hosted LLMs, which would not be feasible in a fully local implementation.

4.3 Method

4.3.1 Fusion network of modular binary experts

We consider the problem of multiclass classification of multimodal time-series observations. Let $m \in \{1, \ldots, M\}$ be an index for modalities (e.g., sensor channels), and let $y \in \{1, \ldots, N\}$ denote the class label with $c \in \{1, \ldots, C\}$ classes. Each example consists of modality-specific time series $(x_1, \ldots, x_M)$, where $x_m \in \mathbb{R}^{T_m \times d_m}$ may have modality-dependent length T_m and dimensionality d_m. The goal is to learn a predictor $\hat{y} = F(x_1, \ldots, x_M)$ that maps multimodal time series to a multiclass label.

To make modality structure and class-conditional evidence explicit, we reduce the multiclass problem to a set of one-vs-rest binary tasks as shown in Figure 4.1. For each modality–class pair

(m, c), we train a binary expert

$$f_{m,c} : x_m \mapsto \hat{p}_{m,c} \in [0, 1],$$

where $\hat{p}_{m,c} \approx p(y = c \mid x_m)$ and the target is $y = c$ versus $y \neq c$. Experts (binary classification models such as recurrent neural networks) are trained independently using a binary classification loss, with class imbalance handled as needed (e.g., class-weighted binary cross-entropy or focal loss). This yields $M \times N$ trained experts $\{f_{m,c}\}$. The baseline expert model (before neural architecture search is performed) is a 1D convolutional residual-Inception network, with a convolutional stem followed by parallel multi-scale temporal branches, normalization, GELU activations, dropout, and global average pooling before a final linear output layer.

Each expert exposes a penultimate representation by removing the final classification layer. We write each expert as

$$f_{m,c}(x_m) = \sigma\big(w_{m,c}^{\top} h_{m,c}(x_m) + b_{m,c}\big)\,,$$

where $h_{m,c}(x_m) \in \mathbb{R}^{D_{m,c}}$ is the penultimate embedding, $(w_{m,c}, b_{m,c})$ are the parameters of the final linear layer, and $\sigma(\cdot)$ is the logistic sigmoid function. In the fusion stage, we discard the final layer and retain $h_{m,c}(x_m)$ as a learned, modality- and class-conditional representation.

We form a joint representation by concatenating all expert embeddings:

$$\begin{aligned} z(x_1, \ldots, x_M) = [\, & h_{1,1}(x_1); \ldots; h_{1,N}(x_1); \\ & \ldots; h_{M,1}(x_M); \ldots; h_{M,N}(x_M)\,]. \end{aligned}$$

Then, a fusion model $g(\cdot)$ (a multilayer perceptron) maps z to multiclass logits $\ell \in \mathbb{R}^N$, followed by a softmax:

$$\ell = g(z), \qquad \hat{y} = \arg\max_c \operatorname{softmax}(\ell)_c.$$

Our default training protocol is modular and performed in two stages: experts are trained first and then frozen, after which only the fusion network is trained. Intuitively, each expert $f_{m,c}$ learns what modality m has to say about whether an example belongs to class c. Because the expert is trained only on the binary question "is this class c or not?" its internal representation is trained to encode class-conditional evidence specific to both the modality

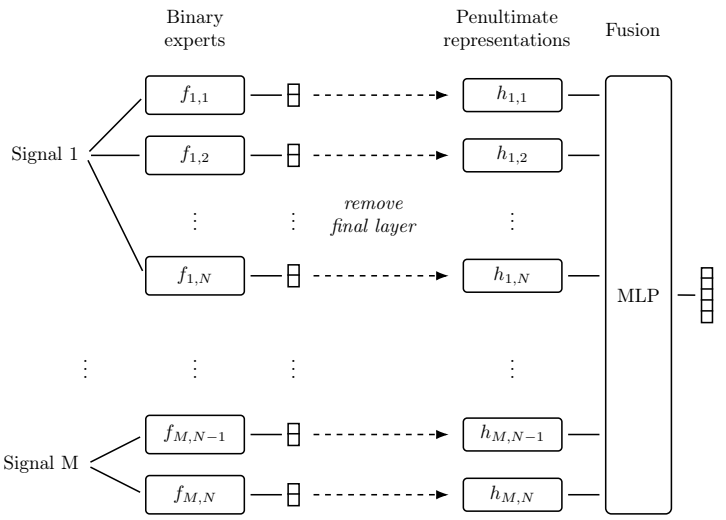


Figure 4.1: Overview of the expert decomposition and representation-level fusion.

and the target class. For example, in sleep-stage classification, an EMG expert for REM may learn patterns associated with reduced muscle tone, whereas an EOG expert for REM may learn to recognize rapid eye movements. The expert embedding $h_{m,c}(x_m)$ can therefore be interpreted as a modality-specific summary of evidence for class c, rather than as a generic representation shared across all classes.

4.3.2 LLM-guided neural architecture search for expert models

Neural architecture search is applied to the individual expert models $\{f_{m,c}\}$. We implement architecture search as an agentic loop in which a remote LLM serves as the *search controller*, proposing candidate expert architectures based on observed fitness trends (validation performance and learning dynamics) accumulated across prior evaluations. All training and evaluation are executed locally under a fixed protocol and recorded as structured artifacts. Figure 4.2 shows the NAS loop.

The search space is intentionally flexible. Rather than selecting from a small hand-designed menu of templates, the remote

controller is allowed to propose substantial changes to the expert pipeline, including preprocessing, feature-extractor structure, layer widths and depths, kernel configurations, normalization, pooling, dropout and other hyperparameters. In practice, the remote LLM proposes architecture and preprocessing code, while the local executor runs all candidates under a fixed evaluation protocol. Thus, the controller has broad freedom at the level of model design, but not at the level of data access.

The controller receives only non-sensitive summaries from the local environment, such as architecture descriptors, hyperparameters, training budgets, validation metrics, optionally learning-curve summaries, runtime, and failure traces. The controller does not access raw data, examples, or any dataset-derived artifacts beyond aggregated metrics. The controller then outputs a new architecture, which is then evaluated locally.

Communication between the remote controller and the on-premise executor is mediated by files in a shared filesystem. Human-readable files such as `manifest.md` preserve the current search context, recent findings, and experiment rationale across controller calls, whereas structured JSON files such as `results.json` define executable instructions and machine-readable outputs. In particular, the controller writes a structured directive specifying which binary experts to target and which candidate preprocessing and model architecture to use. The local executor validates this directive, loads the corresponding candidate code, performs preprocessing, trains and evaluates the requested experts, and writes back structured results containing validation metrics, learning-curve summaries, runtime, and failure traces. This artifact-based interface makes the search loop restartable and auditable: every proposal, execution, outcome, and repair attempt is preserved in the experimental record. It also ensures that the remote LLM never has access to the dataset.

Given a proposed architecture for an expert $f_{m,c}$, the local executor trains the corresponding binary classifier under a fixed training protocol and computes fitness on a held-out validation split using an appropriate binary metric (e.g., AUC, F_1, or balanced accuracy). The evaluation protocol (data splits, loss definition, optimization settings, and training budget) is fixed for the duration of a search run to ensure comparability across proposals. The resulting fitness and auxiliary summaries are appended to the search history and returned to the controller.

The overall procedure is orchestrated as a persistent research cycle that repeatedly (1) proposes expert architecture candidates, (2) instantiates and trains candidates locally, (3) evaluates and ranks results, and (4) records all relevant artifacts. This design supports post-hoc auditing since the full sequence of proposals and outcomes can be reconstructed from the artifact trail.

In deployments with data-locality constraints, dataset access and model training remain strictly on-premise. The remote controller interacts with the local environment only through the constrained artifact interface described above, ensuring that sensitive data are not exposed through prompts or logs while still enabling controller-driven adaptation based on fitness trends.

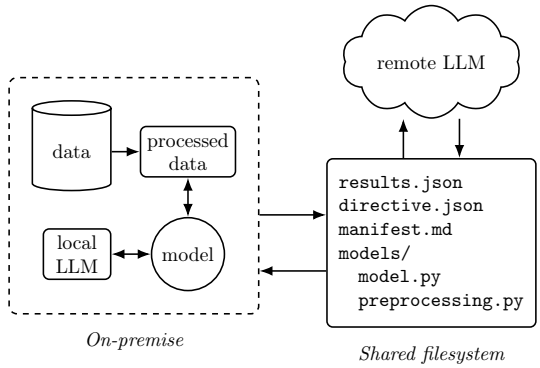


Figure 4.2: Schematic of the data-local autonomous NAS loop. The data and all training remain on-premise. The local execution (model debugging, training and hyperparameter optimization) consumes directives and writes results to a shared filesystem. A remote LLM interacts only through these non-sensitive files. A smaller, local LLM debugs simple errors in the code proposed by the remote LLM, since error messages may contain sensitive data.

4.3.3 Experimental setup and benchmarks

We evaluate our method on two regimes. Firstly, we use the UEA30 dataset, a widely used benchmark for multivariate, multiclass time-series classification [180]. Details about the datasets are reported in Table 4.1. Secondly, we use the SleepEDFx dataset, a sleep staging benchmark dataset with heterogeneous clinical modalities such as EEG, EOG and EMG. This data set is available through

Physionet and consists of polysomnography recordings of healthy adults [181].

We evaluated three configurations:

1. **Baseline end-to-end:** A baseline trained directly on the multiclass task in an end-to-end approach,
2. **Baseline staged:** A baseline in which binary experts were trained individually and then fused as described in section 4.3.1, and
3. **LLM NAS:** the proposed staged system with LLM-guided NAS for a binary expert model architecture.

For each dataset, train/validation/test splits were fixed for the duration of an experiment. For comparison to prior work, we report a reference range of the accuracy of methods from the literature. For the UEA30 datasets, this range was constructed as the minimum and maximum test accuracy reported across three recent benchmark studies spanning 2021–2025 [180], [182], [183]. We selected studies that covered the UEA archive broadly, rather than papers reporting results on only one or a few datasets, so that the comparison would reflect methods evaluated across the benchmark suite rather than dataset-specific tuning. For SleepEDFx, the reference range was constructed from recent studies reporting deep learning 5-class sleep staging on SleepEDFx [45], [184]–[186]. The implementation of our methods can be found at https://github.com/emilhar/arl.

4.4 Results

4.4.1 Experimental overview

Table 4.1 summarizes dataset characteristics and test set accuracy for three configurations: (1) the baseline model trained end-to-end, (2) the baseline model trained in stages, with the binary experts trained individually and then fused, and (3) the LLM-guided NAS model. Across the benchmarks and tested configurations, we make two broad findings: First, the staged expert model is already a strong baseline and often improves upon direct end-to-end training. Second, the LLM-guided NAS can further improve the baseline, although gains depend on the dataset. Unless otherwise

stated, the NAS was run with a 10-cycle budget. Longer searches were performed on two illustrative cases: SleepEDFx, where the NAS method is evaluated on a multimodal clinical problem, and EthanolConcentration, where the baseline model leaves substantial room for improvement.

Table 4.1: Benchmark overview and performance summary. For each dataset, the table reports the number of modalities (m), number of variates (d_m), sequence length (T_m), number of classes (C), train and test set sizes, and test accuracy (%) for three configurations: the end-to-end baseline, the staged baseline with individually trained binary experts followed by fusion, and the proposed LLM-guided NAS model. The final column reports the range of accuracies reported in recent published work for the respective dataset. Unless otherwise stated, NAS was run with a budget of 10 cycles. "–" indicates that the configuration was not evaluated, either because the dataset was too small or because the baseline model achieved 100% accuracy on the validation set.

Dataset	Mod, m	Dim, d_m	Len, T_m	Class, C	Train size	Test size	Baseline end-to-end	Baseline staged	LLM NAS	Previous work
SEDF	3	5	3000	5	112,287	10,774	82.9	84.7	87.9[a]	[82.8–94.0]
AWR	1	9	144	25	275	300	91.7	96.3	–[b]	[94.3–99.6]
AF	1	2	640	3	15	15	33.3	33.3	–	[6.7–74.0]
BM	2	6	100	4	40	40	100	92.5	–[b]	[72.2–100.0]
CT	1	3	0	20	1422	1436	90.1	95.8	95.8	[98.7–99.6]
CR	1	6	1197	12	108	72	95.8	93.1	–[b]	[93.1–100.0]
ER	1	4	65	6	30	270	90.7	85.2	80.1	[82.4–98.9]
EW	1	6	17984	5	128	131	75.6	77.8	84.0	[42.0–95.4]
EP	1	3	206	4	137	138	96.4	97.1	–[b]	[67.0–100.0]
EC	1	3	1751	4	261	263	28.1	32.7	44.1[c]	[27.8–82.4]
HMD	1	10	400	4	160	74	40.5	20.3	21.6	[21.6–59.5]
HW	1	3	152	26	150	850	13.1	14.0	14.0	[18.0–65.7]
JV	1	12	29	9	270	370	96.5	97.0	94.9	[91.4–98.9]
LIB	1	2	45	15	180	180	60.6	76.7	78.9	[75.6–94.1]
LSST	1	6	36	14	2459	2466	49.1	66.5	66.5	[34.0–66.4]
NAT	1	24	51	6	180	180	88.9	84.7	90.1	[76.1–97.1]

Continued on next page

Table 4.1: Benchmark overview for the LLM-guided NAS, continued.

Dataset	Mod, m	Dim, d_m	Len, T_m	Class, C	Train size	Test size	Baseline end-to-end	Baseline staged	LLM NAS	Previous work
PD	1	2	8	10	7494	3498	96.2	95.7	95.7	[87.5–99.7]
PS	1	11	217	39	3315	3353	23.6	26.2	26.2	[10.2–36.7]
RS	2	6	30	4	151	152	79.6	75.7	73.0	[73.7–92.8]
SAD	1	13	93	10	6599	2199	95.2	98.5	98.5	[98.3–99.5]
SWJ	1	4	2500	3	12	15	33.3	33.3	–	[22.0–66.7]
UWG	1	3	315	8	120	320	49.7	83.1	83.1	[71.1–94.4]

[a] 30 cycle search.
[b] NAS was automatically skipped because the validation set showed 100% accuracy.
[c] 60 cycle search.
Abbreviations: SEDF (SleepEDFx), AWR (ArticularyWordRecognition), AF (AtrialFibrillation), BM (BasicMotions), CT (CharacterTrajectories), CR (Cricket), ER (ERing), EW (EigenWorms), EP (Epilepsy), EC (EthanolConcentration), HMD (HandMovementDirection), HW (Handwriting), JV (JapaneseVowels), LIB (Libras), LSST (LSST), NAT (NATOPS), PD (PenDigits), PS (PhonemeSpectra), RS (RacketSports), SAD (SpokenArabicDigits), SWJ (StandWalkJump), and UWG (UWaveGestureLibrary).

4.4.2 Baseline model

The staged baseline performs competitively before LLM-guided NAS is applied as can be seen in Table 4.1. On SleepEDFx, the staged baseline model reaches 84.7% accuracy, falling within the range reported in prior work (82.8-94.0). As seen across the UEA datasets, training the baseline model in stages improves on the end-to-end baseline in 12 of 21 cases, suggesting that training the binary experts individually before fusion is often, but not uniformly, beneficial.

4.4.3 Effect of LLM-guided NAS

The effect of the LLM-guided NAS is mixed across datasets. The clearest gain, relative to the range of previous work, is observed in SleepEDFx, where a 30-cycle NAS increases the test accuracy from 84.7% to 87.9%. This moves the model well into the state-of-the-art range found in recently published results and beyond inter-scorer variability among human sleep scorers [26]. This shows that the LLM-guided NAS can make meaningful improvements to an already competitive baseline model. A second case where the benefits of the LLM-guided NAS are apparent is EthanolConcentration (EC) dataset, where a longer, 60-cycle search improves accuracy from 32% to 44.1%. In contrast to SleepEDFx, in this dataset, the search starts from a relatively weak baseline, suggesting that the search procedure can recover performance even when the initial expert architecture is poorly matched to the task. Looking at Table 4.1, the effect of the LLM-guided NAS is dependent on the specific dataset.

More modest improvements are observed on several other datasets, including EigenWorms (EW, 77.8% to 84.0%), Libras (LIB, 76.7% to 78.9%), and NAT (NATOPS, 84.7% to 90.1%). However, several datasets show no improvement, while a few degrade slightly, likely due to overfitting. This suggests that the LLM-guided NAS is most useful when the baseline model architecture is sub-optimal for the dataset.

Figures 4.3 and 4.4 show the search dynamics of the LLM-guided NAS. Each panel corresponds to one class-dimension expert. Circles show the F1 score obtained when that expert was trained in a given cycle, while dashed lines show the best F_1 achieved so far for that expert. Because only one class-modality architecture is proposed per cycle, many panels show no change for extended

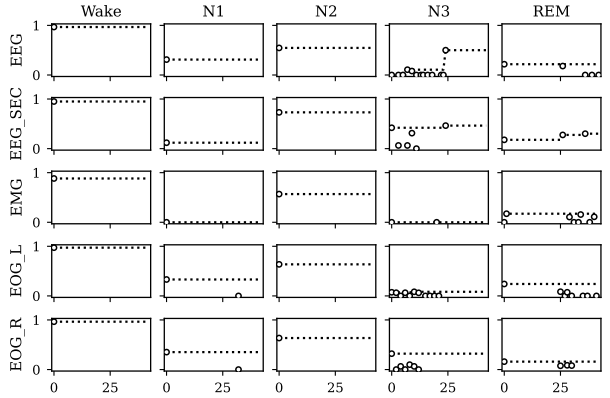


Figure 4.3: Search trajectories for binary experts during a 30-cycle NAS on the SleepEDFx dataset. Each panel corresponds to one class-modality expert. Dashed lines show the F_1 score of the best found expert, and circles show the F_1 score of the expert trained during the cycle.

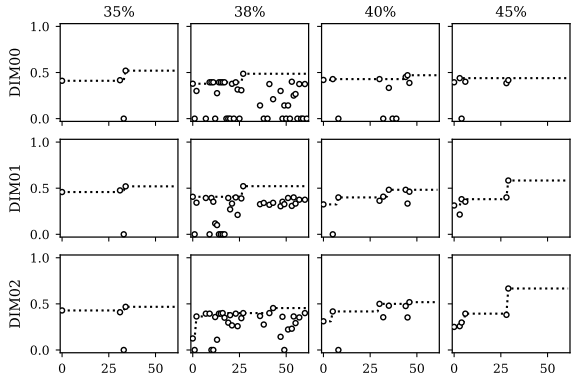


Figure 4.4: Search trajectories for binary experts during a 60-cycle NAS on the EthanolConcentration dataset. Each panel corresponds to one class-modality expert. Dashed lines show the F_1 score of the best found expert, and circles show the F_1 score of the expert trained during the cycle.

periods, and flat lines indicate either that the expert was not examined or that retraining did not improve on the current best result. Figure 4.3 shows that on SleepEDFx, gains are concentrated on the weakest experts (N3 and REM sleep), whereas in Figure 4.4 we see improvements across all experts.

4.4.4 Overall takeaways

These results support three main conclusions. First, the staged expert-fusion baseline model is already a strong model, fit for a broad set of multivariate multiclass time-series classification tasks. Second, the LLM-guided NAS can provide additional gains, with the largest improvements in cases where the baseline architecture is a poor fit or where complex, multimodal data leaves room for targeted expert adaptation. Third, the practical value of the proposed framework extends beyond benchmark accuracy alone. Instead, it lies in enabling automated, long-horizon architecture search in sensitive, practical situations, where raw data cannot be exposed to an external LLM.

4.5 Discussion

4.5.1 Open Access Tool

The proposed framework is accessible as an open-source software repository under the MIT license. Our development and open-access sharing of this tool enable ML researchers to systematically and automatically design models for multimodal time-series classification. Furthermore, the tool's design is sufficiently interpretable, allowing domain experts with no in-depth ML knowledge to apply these methods. A core strength of our method is that the user can understand what modality and class is being worked on and see the reasoning behind the LLM's model suggestion. In the case of sleep staging, for example, during one cycle of our tests the system reported that REM was the weakest ensemble class and that improving it would lift macro-F1 the most. The LLM controller further pointed out that EMG expert had zero F1 for REM, even though REM is characterized by low EMG tone and atonia, and consequently proposed that simple amplitude features could unlock strong precision gains. Since this information is shared in a human-readable way, a domain expert can monitor the process

and use their domain knowledge to control the model development, without requiring in-depth ML knowledge to implement the changes.

On the one hand, this more intuitive human approach allows people with in-depth knowledge of the data to apply ML, which may lead to new perspectives and contributions. Our results are in line with Miguel-Morante et al., who argue that this human interpretability adds value to the system [176]. On the other hand, there is a risk associated with making ML accessible to people without in-depth ML knowledge [170]. This could lead to applications that are not properly validated, overfitting, or being misinterpreted, potentially resulting in unreliable or clinically misleading outcomes. As Somer points out in their review from 2025, there are ways to mitigate the risk of domain experts using simplified tools to create predictive models, such as setting up an accountability framework, offering special training to domain experts, or increasing the explainability of the model [187].

4.5.2 Data Privacy

This chapter showed that developing a model that stores data and trains locally can still achieve performance comparable to or better than that of common benchmarking datasets for multimodal time series classification. Our system architecture, therefore, adheres to the concept of privacy by design (PBD), which centers on integrating privacy-preserving design into a system's architecture rather than adding privacy measures afterward [188]. Our architecture furthermore strongly aligns with the PBD design principles defined by Cavoukian et al. [189] by i.) enacting privacy proactively, ii.) preserving data privacy by default, and iii.) preserving full functionality of the system.

4.5.3 Limitations

A practical evaluation challenge arises when LLMs are used to assist with planning, debugging, and documenting experiments. Canonical NAS benchmarks and leaderboards are heavily documented, and an assistant model may reproduce benchmark-specific architectures and recipes that are widely available in the public literature and online. This aligns with the work of Tornede et al., who note that the inability to trace which data the LLM used for

training is one of the most severe risks in using and evaluating LLMs for AutoML [190]. This makes it difficult to disentangle performance gains attributable to the proposed search-and-evaluate mechanism from gains due to benchmark-specific prior art. For this reason, we emphasize *domain-representative* multiclass multimodal time-series tasks under data-local constraints, where the objective is robust, auditable experimental progress rather than leader board ranking.

4.6 Conclusion

In this chapter, we presented a novel framework for a data-local autonomous NAS guided by an LLM to tackle the challenging problem of multiclass multimodal time-series classification. Firstly, we turned the classification problem into a fusion network of binary expert classifiers, each of whom trains on a single modality. This allows the model to focus on the properties of each data stream in isolation. The output of the experts is then fused together to form the final classifier. As our results on 22 benchmark datasets demonstrate, this modification already brings significant performance gains. Secondly, we implemented an LLM-guided NAS that had the objective of improving the experts by proposing changes to their model architecture. The NAS trained over multiple cycles and was able to improve the architecture for several of the benchmark datasets, leading to an even greater enhancement in performance in some cases. This LLM-guided NAS, while utilizing the power of modern LLMs, is data privacy preserving, as the models are trained locally, making this a feasible approach for developing ML models in domains such as healthcare.

Although our experiments are limited, our work shows great potential and opens several doors for future work. On the one hand, the fusion network of binary experts could be adapted to regression problems and forecasting, which are common when working with time series, making this modification highly relevant. In this work, we focus on multivariate time series, but other data modalities, such as images and text, could be included as well to make the framework more generalized. On the other hand, the LLM-guided NAS itself is a novel approach to finding optimal architectures in neural networks, and our prototype is a simple one. With more sophisticated LLMs and prompting, we will focus

our future research on faster and more effective learning.

Chapter 5

Staging by the book

Automatic sleep stage classification using scoring rules

Full reference: (Submitted to Journal of Sleep Research.) E. Hardarson *et al.*, 2026, *Staging by the Book: Automatic Sleep Stage Classification Using Scoring Rules*, DOI: 10.48550/ARXIV.2605.22859

5.1 Introduction

Many decisions in clinical practice are governed by written rules and guidelines. When such decisions are automated, the goal is not only to produce an accurate output, but also to preserve the logic by which the decision is supposed to be made under the rules [191], [192]. Automation in clinical decision-making is increasingly performed with supervised Machine Learning (ML) algorithms, which learn statistical patterns that produce high agreement with human labels, without being explicitly constrained to the rules governing the decision [136], [138], [191].

Sleep stage scoring is an example of such a rule-governed clinical decision-making workflow. It is a foundational step in sleep research and clinical practice, serving as the basis for diagnosing sleep disorders and understanding the physiology of sleep. The gold standard for measuring sleep is a sleep study called polysomnog-

raphy (PSG), which involves the overnight recording of multiple physiological signals, such as electroencephalography (EEG), electrooculography (EOG) and chin electromyography (EMG) [8].

Clinical sleep stage scoring guidelines, such as those published by the American Academy of Sleep Medicine (AASM), specify that sleep should be staged in consecutive 30 second epochs. Each epoch is classified as one of five sleep stages: Wake, Rapid Eye Movement (REM) sleep (sleep stage R) or Non-REM sleep 1, 2 or 3 (sleep stages N1, N2 or N3). Sleep stage scoring is performed manually by a human sleep expert, in a procedure that typically takes between 1 and 3 hours for each full night recording [11], [12]. Manual sleep stage scoring is subject to inter-scorer variability and agreement differs between sleep stages. Multi-scorer studies have found relatively high agreement for sleep stages Wake, N2 and REM, but lower agreement for sleep stages N1 and N3 [17], [193], [194]. Complete agreement across expert scorers occurs in only a minority of epochs [195], [196]. The growing number of sleep studies [197] and the labor-intensive nature of manual sleep stage scoring have motivated the development of practical and reliable automatic sleep staging algorithms [26].

In recent years, supervised ML, and specifically Deep Learning (DL), has become the dominant methodology for automated sleep staging [19], [42], [198], [199]. ML models trained on large datasets of human-labeled sleep recordings have achieved performance comparable to human inter-scorer agreement [200], [201]. The ML models do this by learning to identify statistical patterns within the signals (see Chapter 4) [29], [202]. However, this training procedure is typically not guaranteed to align with the logic of clinical scoring rules [15]. The reasoning behind the models' predictions is difficult to interpret and explain, a phenomenon termed the "black-box" problem [136].

This opacity of sleep staging ML model reasoning presents a barrier to their clinical adoption [15]. When a human sleep expert scores a sleep study, they can justify their decision by pointing to features defined in clinical guidelines, such as sleep spindles or alpha rhythm. Although ML models may learn to recognize such patterns, they do so by mapping high-dimensional data through millions of internal parameters, rather than through explicit logical steps.

Sleep stage scoring is, by design, a rule-based decision process whose underlying logic is already specified. The AASM scoring

manual provides a deterministic set of rules, so an algorithm does not, in principle, need to infer this logic from training data. Despite this, there is a notable gap in contemporary literature regarding methods that explicitly use the scoring logic described in clinical guidelines.

Historically, automated sleep staging did not begin with machine learning. Early methods attempted to translate human scoring practice into algorithms, first under the Rechtschaffen and Kales (R&K) criteria [14] and later under the first version of the AASM scoring manual [13], [20], [21]. Earlier rule-based methods such as that of [25] have shown that sleep scoring logic could be operationalized with high agreement to human consensus [21]–[24]. The field of automated sleep staging later shifted towards ML and DL approaches because of their strong sleep staging accuracy [26], [27], but rule-based approaches have not disappeared [49].

A practical challenge in operationalizing sleep scoring rules is that the sleep stage label depends on intermediate signal features described in the scoring manual. For example, scoring decisions may depend on whether an epoch contains sleep spindles, K-complexes, slow wave activity, rapid eye movements, alpha rhythm, low-amplitude mixed-frequency activity, or changes in chin EMG tone. In this chapter, we refer to these signal segments that are relevant to sleep stage scoring decisions as *micro-annotations*. Earlier computer-assisted and rule-based sleep staging systems have implicitly operationalized parts of clinical scoring logic, for example through expert-defined features or rules about sleep stage transitions [21]–[23], [25]. However, most contemporary automatic sleep staging systems report sleep stage labels or probabilities, rather than exposing the rule-application process that connects signal features to scoring decisions.

Beyond assigning a sleep stage label, an automated scoring method can also support interpretation by showing how the decision was reached. In sleep staging, this requires connecting signal features, scoring criteria, and the final stage assignment in a form that is familiar to human scorers (see Chapter 9). Few automatic sleep staging systems provide epoch-level natural-language justifications that trace the decision through the scoring rules.

In this chapter, we build on the rule-based tradition of automatic sleep staging by making the operationalization explicit: we translate the logic of the AASM scoring manual into executable code and record the reasoning process behind each sleep stage

assignment. The result is a deterministic and inspectable method in which sleep stage assignments can be traced to the scoring rules and signal annotations used by the algorithm.

In this chapter, we ask: To what extent can the logic of the AASM sleep stage scoring manual be operationalized as deterministic, inspectable code, and what limitations emerge when such a method is evaluated against multi-scorer human consensus?

We address this research question by developing a sequential, logic-driven sleep staging algorithm in Python that operationalizes the logic of the AASM scoring rules as executable code. In addition, we implement an explanation mechanism that records the rule-based reasoning behind each sleep stage assignment and renders this reasoning as natural-language justifications linked to scoring rules.

We evaluated the method using 50 full-night type II PSG recordings scored independently by 10 human sleep experts. The evaluation showed that the method achieved moderate agreement with the multi-scorer consensus, with strongest performance for sleep stages N2 and R. A reviewer-based discrepancy assessment suggested that reviewed disagreements were primarily associated with incomplete or inaccurate micro-annotations, rather than the rule logic itself. While previous rule-based sleep staging methods have used rules as a guideline or processing step, we present what is, to our knowledge, the first method that attempts to explicitly code the clinical logic of the AASM scoring manual into an algorithm [19], [203].

5.2 Method

5.2.1 Algorithm

We designed the automatic sleep staging algorithm to emulate the logic of the AASM scoring manual. The method, written in Python 3.12, determines sleep stages by processing PSG data in three steps as shown in Figure 5.1.

The process begins by loading the raw PSG data stored in European Data Format (EDF), a standard file format for physiological time-series recordings, using the MNE Python library [204], dividing the signals into 30-second epochs, and checking signal integrity. After this, the sleep staging process begins. First, the method establishes attributes pertaining to the full night's

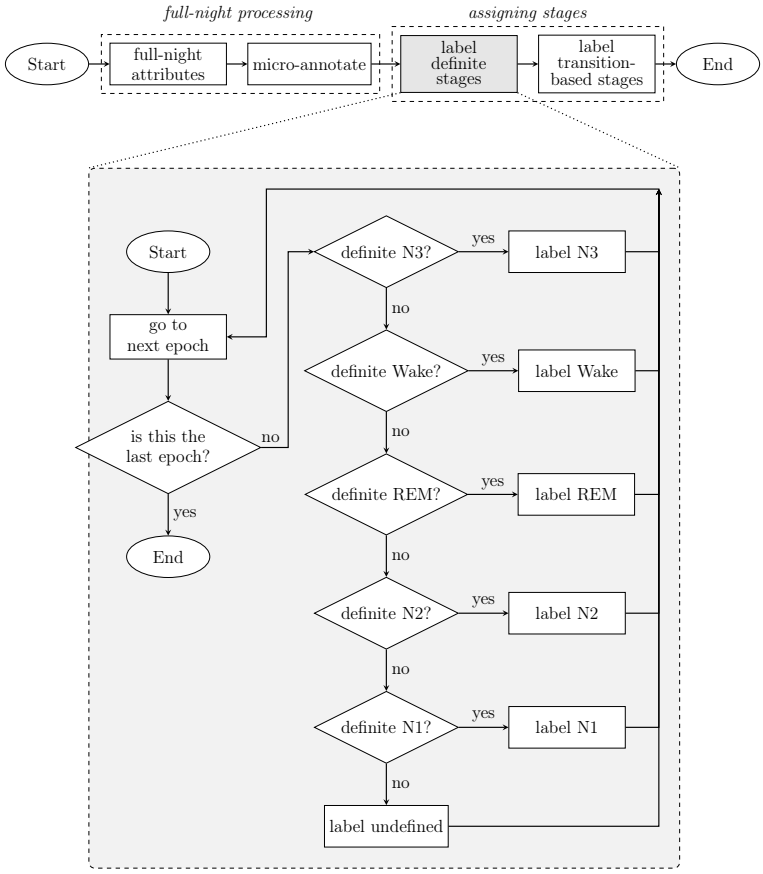


Figure 5.1: Schematic overview of the automatic sleep staging process. The top level shows the high-level steps from initial processing of full-night attributes and micro-annotating to the two-pass sleep staging process. Full-night attributes refer to recording-level information, such as alpha-rhythm status and chin electromyography baseline, that is needed before applying some epoch-level scoring rules. The labeling of definite sleep stages is shown in an expanded view, where each epoch is sequentially compared to criteria for each of the sleep stages. Abbreviations: N1, non-rapid eye movement sleep stage 1; N2, non-rapid eye movement sleep stage 2; N3, non-rapid eye movement sleep stage 3; REM, rapid eye movement sleep.

recording. These *full-night attributes* act as global variables that govern the applicability of certain scoring rules. For example, the boolean variable `generates_alpha_rhythm` is decided via a sliding window Welch periodogram analysis on occipital EEG channels, which estimates the spectral power of the signal over time and identifies whether alpha-band activity is present [8], [205]. This variable distinguishes between the use of the AASM scoring manual's *F2 rule* (sleep should begin when alpha activity is attenuated and replaced by low amplitude mixed frequency EEG signal in patients who generate alpha rhythm) and *F3 rule* (sleep starts at the earliest occurrence of background frequency slowing, vertex sharp waves, or slow eye movements). Another full-night attribute is the chin EMG baseline, which establishes global statistics for the chin EMG signal in order to define thresholds for high and low muscle tone, which is used when classifying sleep stage R.

Second, the algorithm identifies various physiological events described in the scoring rules. These micro-annotations are stored as tuples containing event labels, temporal boundaries and channel derivations. The YASA Python library [201] was used to detect sleep spindles, rapid eye movements (REMs) and slow wave activity (SWA) on frontal and central EEG channels. Alpha rhythm and low amplitude mixed frequency (LAMF) segments were operationalized using spectral features estimated with Welch periodograms [205]. Specifically, power spectral density (PSD) was estimated in sliding windows, and each window was characterized by the relative power in frequency bands relevant to the AASM scoring rules [8]. Windows meeting the predefined alpha or LAMF criteria were marked as corresponding micro-annotations. Alpha rhythm was detected using 2.0-second windows advanced in 0.5-second steps, corresponding to 75% overlap. A window was marked as alpha rhythm when power in the 8–12 Hz band accounted for at least 50% of total spectral power. LAMF was defined heuristically as a contiguous low-amplitude EEG segment lasting at least 1.0 second, where low amplitude was defined as absolute signal amplitude below the channel-level mean absolute amplitude minus 0.01 standard deviations. Candidate low-amplitude segments were then retained as LAMF if power in the 4–7 Hz band accounted for at least 1% of total spectral power.

The algorithm processes the annotated epoch in two sequential passes to assign sleep stages, as shown in Figure 5.1. The first pass evaluates epochs according to *definite* scoring rules in the

same order of precedence as the AASM scoring rules: sleep stage N3, then wake, R, N2 and finally N1, stopping at the first satisfied rule. Sleep stage N3 is assigned if SWA covers more than 20% of the epoch. Only if that criterion is not fulfilled does the method check for evidence of the next sleep stage. Wake is assigned if alpha rhythm or other markers such as eye blinks exceed 50% of the epoch. If not, the epoch is labeled sleep stage R if it contains LAMF and REMs without K-complexes or spindles, provided that EMG tone is low. Sleep stage N2 is assigned if a K-complex unassociated with arousal or a spindle occurs in the first half of the current epoch or the last half of the previous one. Finally, sleep stage N1 is assigned using a simplified alpha-rhythm rule. In alpha-generating recordings, the method assigns N1 when the preceding epoch was Wake, at least 50% of the current epoch consists of LAMF, and alpha rhythm is detected in either the current or preceding epoch.

Epochs that remain undefined after the first pass are re-evaluated in a second pass using *transition rules*, which depend on the context of adjacent epochs. For example, the method continues to score sleep stage R if the previous epoch was sleep stage R and the current epoch exhibits LAMF and low EMG without intervening arousals or spindles.

Epochs which remain undefined after both passes can either inherit the sleep stage of the preceding epoch, or be assigned a sleep stage based on the YASA ML sleep stage classifier, depending on the user's preference. We report results for the inheritance-only configuration. Algorithm 1 shows the step-by-step procedure for translating physiological micro-annotations into sleep stage labels in broad terms.

5.2.2 Explanation generation

One of the strengths of a rule-based sleep staging algorithm is the possibility of providing transparent justifications for every staging decision. The method does this by populating an *explanation log* as it traverses the logical steps. The algorithm appends the result of every sleep stage evaluation to a list within the explanation log, which is implemented as a nested dictionary. Each entry in the list is a tuple containing the sleep stage being evaluated, a boolean indicating if the criteria were met, and a text string detailing the findings. This creates a complete trace of the step-by-step decision

Algorithm 1 Rule-Based Sleep Staging

Require: Epochs $\{E_t\}_{t=1}^{T}$ with micro-annotations.
Ensure: Sleep stage labels y_t and explanations $\mathcal{E}_t$.
Initialize: Set all labels $y_t \leftarrow$ undefined

Pass 1: Definite Rules (Local Epoch Criteria) for *each epoch* E_t *from* 1 *to* T **do**

end
1. Stage N3 Check: If Slow Waves $\geq 20\%$ of epoch $\implies y_t \leftarrow$ N3
2. Wake Check: If Wake-signs (Alpha, Blinks, etc.) $> 50\%$ $\implies y_t \leftarrow$ Wake
3. Stage R Check: If (LAMF + REMs + Low EMG) and No (Spindle/K-Complex) $\implies y_t \leftarrow$ REM
4. Stage N2 Check: If Spindle or K-Complex in current 1st half or previous 2nd half $\implies y_t \leftarrow$ N2
5. Stage N1 Check: If (Alpha attenuated) and (LAMF $> 50\%$) $\implies y_t \leftarrow$ N1
Log Reasoning: Update $\mathcal{E}_t$ with triggered rule and reasons for rejecting other rules.

Pass 2: Transition Rules (Temporal Context) for *each epoch* E_t *where* y_t *is still undefined* **do**

end
REM Continuity: If $y_{t-1} =$ stage R and features remain consistent $\implies y_t \leftarrow$ stage R
N2 Continuity: If $y_{t-1} =$ stage N2 and no evidence of stage change $\implies y_t \leftarrow$ stage N2
Inheritance: If still undefined, inherit stage from y_{t-1}.
Log Reasoning: Document the use of preceding epoch in $\mathcal{E}_t$.

return Sleep stage labels and explanations $\{(y_t, \mathcal{E}_t)\}_{t=1}^{T}$

process.

The explanation log shows how the final sleep stage is reached by sequentially ruling out alternative candidate sleep stages until the final decision is reached. Because the algorithm stops searching once a definite sleep stage is assigned, the logs naturally present a chronological account of what was tried, why certain sleep stages were rejected (e.g., "slow wave activity at 15%, below the 20% threshold"), and which rule ultimately triggered the label.

The explanation log is a technical data structure designed for programmatic tracking of operations. To facilitate the user's deeper understanding of the logical reasoning behind staging decisions, the explanation log is fed into a Large Language Model (LLM;specifically OpenAI GPT 4). The LLM's system prompt is constructed with the epoch's explanation log and the Python source code used for the classification and explanation log. This provides the LLM with the context necessary to interpret the staging decision and justify them by referencing specific events in the signals, as well as the AASM scoring rules.

5.2.3 Dataset

We evaluated our method on the Sleep Revolution Multicenter Scoring dataset which consists of 50 prospective full-night type II PSG recordings collected at Reykjavik University (February - June 2021) using a Nox A1 system [17]. Each recording was independently scored by 10 human sleep experts from seven sleep centers (April - September 2021) in Europe and Australia following AASM v2.6 criteria, enabling both a majority-vote consensus hypnogram and a per-epoch agreement ratio (fraction of scorers assigning the most frequently scored sleep stage).

The cohort is adult and mixed with respect to common sleep disorders (healthy, obstructive sleep apnea, insomnia and restless legs syndrome symptoms) (58% male, mean age 42.9±13.7 years, mean BMI 27.3±5.8 kg/m^2 and mean AHI 15.2±15.6 events/h) [17]. One PSG was used in the development of the method for manual iterative rule implementation and parameter tuning and was subsequently excluded from the primary evaluation. The remaining 49 PSGs form the test set.

5.2.4 Evaluation

The performance of the method was evaluated using the dataset, with the majority vote consensus hypnogram of the ten scorers as a reference. We assessed agreement both in terms of the epoch-by-epoch sleep stage classification and sleep architecture metrics. We calculated overall staging accuracy and Cohen's kappa, as well as confusion matrices, to identify sleep-stage-specific patterns of disagreement. Before calculating agreement and sleep-architecture metrics, each recording was cropped to an analysis period. Following Nikkonen *et al.* [17], this period was defined as the interval from the first epoch to the last epoch in which at least one human scorer assigned a sleep stage other than Wake. Leading and trailing all-Wake periods were removed, while Wake epochs within this interval were retained. The same cropping procedure was applied to the development recording and to all test recordings.

To complement this performance evaluation, a reviewer with sleep scoring experience conducted a reviewer-based discrepancy assessment to characterize sources of disagreement between the method and the reference. This assessment targeted two categories: (1) Epochs characterized by high human agreement where the method did not match the consensus, and (2) epochs with low human agreement where the method matched the sleep stage of the majority-vote consensus. A selection of epochs from both categories were investigated to determine whether the disagreement was associated with inaccurate or incomplete micro-annotations, rule application, or another source. By examining epochs with high human disagreement, our analysis sought to determine if the method's classification could provide insight into the reasons for ambiguity in human scoring.

We randomly sampled 20 epochs from the PSG used for method development, which had been used during iterative rule implementation and parameter tuning. For each sampled epoch, the reviewer first assigned a sleep stage using the PSG signals and the AASM scoring manual, without knowing the sleep stage label from the human sleep experts and the sleep staging algorithm. Next, the reviewer was shown the human sleep expert consensus sleep stage label and the agreement ratio, followed by the sleep stage label predicted by the algorithm and the algorithm's explanation. The reviewer then answered three yes/no questions: (1) Were the micro-annotations used by the algorithm correct? (2) Given those

micro-annotations, was the applied scoring rule correct? (3) Was the assignment of definite or transition sleep stage label correct? If the reviewer determined that micro-annotations were incorrect, they could answer "not applicable" to the other two if deemed necessary.

5.3 Results

5.3.1 Output data

For each PSG recording, the method outputs a hypnogram consisting of epoch-by-epoch sleep stages, along with an epoch-by-epoch visualization of the micro-annotations. Additionally, the method produces the explanation log for every staging decision. Figure 5.2 shows an example of an epoch which the algorithm classified as N2. The figure shows the signals used, along with the micro-annotations.

Along with inspecting the raw explanation log, the user has two ways to view explanations: *static-* and *interactive* explanations. In static explanations, the log is rendered as a list of elimination rules that outline why each candidate sleep stage was not assigned until the final rule triggers the selected label, in the order which they were applied. Figure 5.3 shows a static explanation generated by the algorithm for the same epoch as shown in Figure 5.2.

Interactive explanations can be explored in a conversation with an LLM through a chat interface. The LLM (OpenAI GPT 4) summarizes each epoch's explanation log in concise natural language and responds to the user's questions. Figure 5.4 shows an example of an interactive explanation.

Given identical inputs, the method is deterministic: repeated runs produce identical hypnograms, micro-annotations and explanation logs. However, the LLM responses in the interactive explanations vary across runs, due to probabilistic sampling during the LLM's sampling process [206].

The average runtime for a full-night recording was approximately 90 minutes on an Intel Core i7-8650U CPU, with the overwhelming majority of computation time spent in preprocessing and annotation.

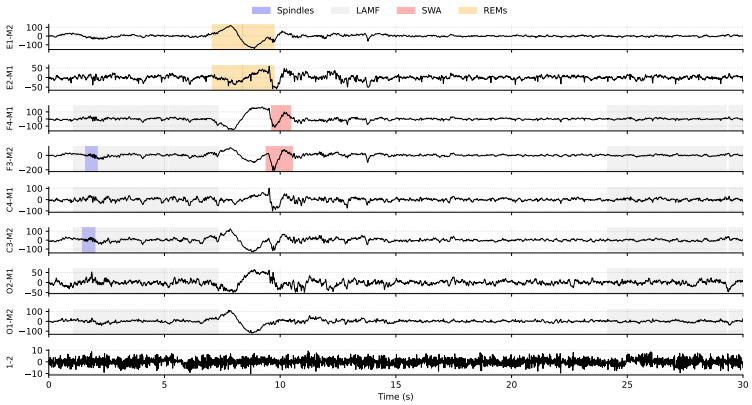


Figure 5.2: One epoch with micro-annotations. The displayed channels include electrooculography (EOG: E1-M2, E2-M1), electroencephalography (EEG: F4-M1, F3-M2, C4-M1, C3-M2, O2-M1, O1-M2), and chin electromyography (EMG: 1-2). Colored overlays indicate detected sleep spindles, low-amplitude mixed-frequency activity (LAMF), slow wave activity (SWA), and rapid eye movements (REMs). This epoch was classified as sleep stage N2. Abbreviations: EEG, electroencephalography; EOG, electrooculography; EMG, electromyography; LAMF, low-amplitude mixed-frequency activity; SWA, slow wave activity; REMs, rapid eye movements; N2, non-rapid eye movement sleep stage 2.

5.3.2 Comparison to human consensus

To evaluate agreement with human scoring, we compared the algorithm's epoch-by-epoch sleep stage labels to the majority-vote consensus hypnogram of the ten human scorers in the dataset. As described in Section 5.2.3, one PSG recording was used during the development of the algorithm to iterate on rule implementation and parameter tuning. This recording was excluded from the test set and we report results on this *development recording* separately.

Tables 5.1 and 5.2 show sleep stage confusion matrices for the development PSG, expressed as row-normalized percentages and total counts, respectively. Tables 5.3 and 5.4 show the same but aggregated across all subjects in the test set.

Overall agreement with majority-vote consensus across the dataset was 60.5% with Cohen's κ=0.42. Agreement was highest for N2 (recall 83.5%) and moderate for sleep stage R (recall 68.7%)

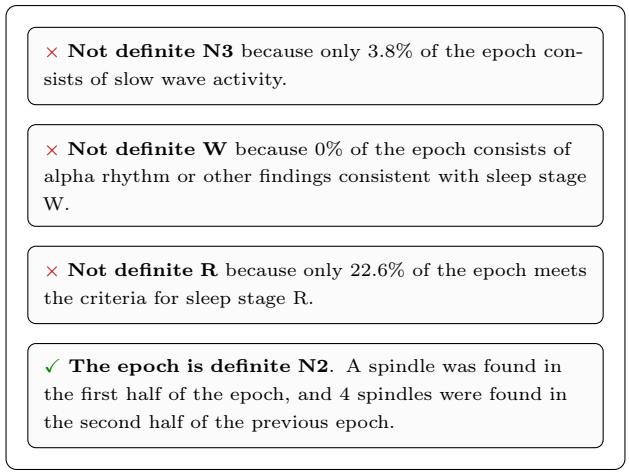


Figure 5.3: Example of a sequential elimination trace produced by the rule-based sleep staging method. The explanation rules out incompatible sleep stages in sequence before stating the final sleep stage assignment and its supporting evidence. Abbreviations: W, Wake; R, rapid eye movement sleep; N2, non-rapid eye movement sleep stage 2; N3, non-rapid eye movement sleep stage 3.

and N3 (recall 50.5%). Wake and N1 recall were low (14.5% and 5.9%, respectively).

Table 5.5 shows the method's mean absolute difference from sleep architecture metrics derived from the majority-vote consensus hypnograms. For comparison, the table also shows the mean absolute difference between individual human scorers and the consensus of the other nine scorers. Results are shown separately for the test set and for the development recording. In the test set, the method showed larger differences from the consensus-derived metrics than individual human scorers across all evaluated sleep metrics in the test set.

5.3.3 Assessment of discrepancies

Figures 5.5 and 5.6 show examples of hypnograms comparing the algorithm's sleep stage predictions to the majority-vote consensus. In both figures, epochs where the algorithm disagrees with the consensus sleep stage are highlighted in red. The consensus

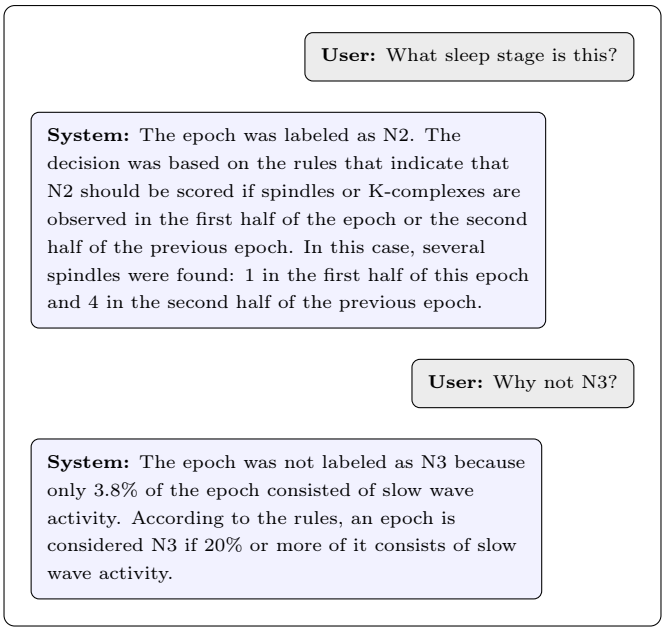


Figure 5.4: Example of a natural-language explanation dialogue produced by the rule-based sleep staging method, showing both the justification for the assigned sleep stage and the rejection of an alternative candidate sleep stage. Abbreviations: N2, non-rapid eye movement sleep stage 2; N3, non-rapid eye movement sleep stage 3.

hypnogram is highlighted in a progressively darker shade of yellow, depending on the level of disagreement among the 10 scorers. Many epochs where the algorithm disagreed with the reference occurred in areas where human agreement was low. This can also be seen in Figure 5.7, which shows the distribution of human agreement ratios for epochs where the algorithm agreed with the reference, compared with epochs where it disagreed.

The reviewer's discrepancy assessment suggested that when the reviewer agreed with the method's sleep stage label, the decision was usually supported by the appropriate reasoning. In 71% of epochs where the reviewer agreed with the method's sleep stage label, the micro-annotations, rule application, and definite/transition assignment were all deemed correct by the reviewer. In epochs where the reviewer disagreed with the method's label, the

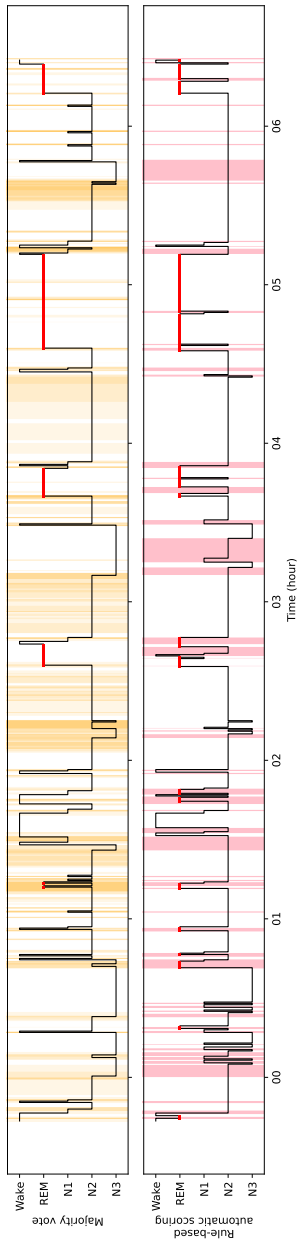


Figure 5.5: The top panel shows the human consensus hypnogram for the polysomnography used during the development of the method, with areas of high human disagreement highlighted. The bottom panel shows a predicted hypnogram for the same polysomnography, with disagreements with the human consensus hypnogram highlighted in red. Abbreviations: N1, non-rapid eye movement sleep stage 1; N2, non-rapid eye movement sleep stage 2; N3, non-rapid eye movement sleep stage 3; REM, rapid eye movement sleep.

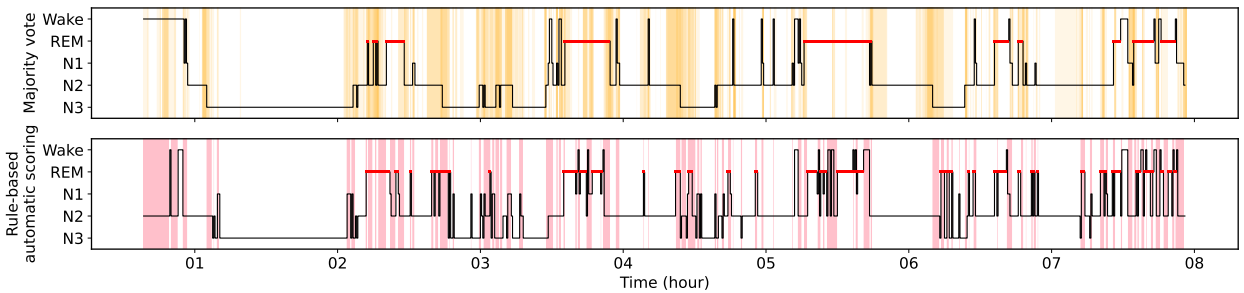


Figure 5.6: The top panel shows the human consensus hypnogram for one of the test polysomnographies with areas of high human disagreement highlighted. The bottom panel shows a predicted hypnogram for the same polysomnography, with disagreements with the human consensus hypnogram highlighted in red. Abbreviations: N1, non-rapid eye movement sleep stage 1; N2, non-rapid eye movement sleep stage 2; N3, non-rapid eye movement sleep stage 3; REM, rapid eye movement sleep.

Table 5.1: Confusion matrix for the single development polysomnography (excluded from the primary evaluation), comparing epoch-by-epoch predictions against the majority-vote consensus of ten human sleep experts, expressed in percentages. Each row is normalized to represent 100% of the epochs for that reference sleep stage. Abbreviations: N1, non-rapid eye movement sleep stage 1; N2, non-rapid eye movement sleep stage 2; N3, non-rapid eye movement sleep stage 3; R, rapid eye movement sleep.

Reference	**Wake**	**N1**	**N2**	**N3**	**R**
Wake	40.7%	3.7%	37.0%	1.9%	16.7%
N1	2.8%	11.1%	63.9%	0.0%	22.2%
N2	0.0%	2.0%	94.8%	0.2%	2.9%
N3	0.0%	10.1%	43.5%	44.2%	2.2%
R	0.8%	3.0%	11.4%	0.0%	84.8%

micro-annotations were judged incorrect in all cases. These findings suggest that, among the reviewed examples, reviewer-method disagreements were primarily associated with micro-annotation quality, rather than the rule logic itself.

Table 5.2: Confusion matrix for the single development polysomnography (excluded from the primary evaluation), expressed in total epoch counts. Abbreviations: N1, non-rapid eye movement sleep stage 1; N2, non-rapid eye movement sleep stage 2; N3, non-rapid eye movement sleep stage 3; R, rapid eye movement sleep.

Reference	**Wake**	**N1**	**N2**	**N3**	**R**
Wake	22	2	20	1	9
N1	1	4	23	0	8
N2	0	9	421	1	13
N3	0	14	60	61	3
R	1	4	15	0	112

Table 5.3: Confusion matrix for all test recordings in the dataset, comparing the algorithm's epoch-by-epoch predictions against the majority-vote consensus of ten human sleep experts, expressed in percentages. Each row is normalized to represent 100% of the epochs for that specific reference sleep stage. The rows show the reference sleep stages, while the columns show the sleep stages assigned by the algorithm. Abbreviations: N1, non-rapid eye movement sleep stage 1; N2, non-rapid eye movement sleep stage 2; N3, non-rapid eye movement sleep stage 3; R, rapid eye movement sleep.

	Method (%)				
Reference	**Wake**	**N1**	**N2**	**N3**	**R**
Wake	14.5	3.7	43.4	3.4	35.0
N1	0.1	5.9	54.7	0.3	39.1
N2	0.0	2.7	83.5	1.7	12.1
N3	0.0	9.2	24.5	50.5	15.8
R	0.4	2.0	28.9	0.1	68.7

Table 5.4: Confusion matrix for all test recordings in the dataset, comparing the algorithm's epoch-by-epoch predictions against the majority-vote consensus of ten human sleep experts, expressed in total epoch counts. The rows show the reference sleep stages, while the columns show the sleep stages assigned by the algorithm. Abbreviations: N1, non-rapid eye movement sleep stage 1; N2, non-rapid eye movement sleep stage 2; N3, non-rapid eye movement sleep stage 3; R, rapid eye movement sleep.

	Method				
Reference	**Wake**	**N1**	**N2**	**N3**	**R**
Wake	895	228	2676	209	2154
N1	2	191	1780	10	1272
N2	7	553	17228	358	2497
N3	0	805	2151	4429	1381
R	36	181	2646	11	6294

Table 5.5: Comparison of method and human mean absolute differences (MAD) from consensus-derived sleep metrics. *Method MAD* and *Human MAD* refer to the test set. *Dev. Method MAD* and *Dev. Human MAD* refer to the single development polysomnography excluded from the primary test-set evaluation.

Metric	**Method MAD**	**Human MAD**	**Dev. Method MAD**	**Dev. Human MAD**
Total Sleep Time (min)	53.3	12.2	15.0	3.8
Sleep Efficiency (%)	10.6	2.4	3.7	0.9
Wake After Sleep Onset (min)	53.3	12.2	15.0	3.8
N1 Duration (min)	19.4	14.9	1.5	7.8
N2 Duration (min)	64.7	27.9	47.5	24.9
N3 Duration (min)	40.5	22.6	37.5	22.4
REM Duration (min)	54.1	9.5	6.5	5.5
N1 proportion (%)	4.6	3.5	0.6	2.1
N2 proportion (%)	10.1	6.5	9.9	6.6
N3 proportion (%)	10.6	5.2	10.3	5.8
REM proportion (%)	10.0	2.4	1.0	1.4

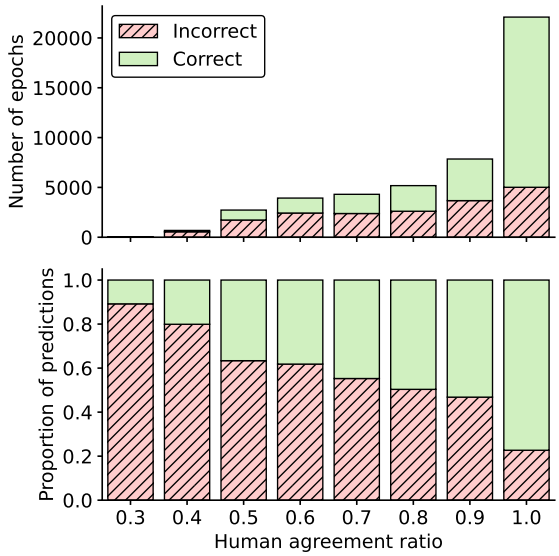


Figure 5.7: Distribution of human inter-scorer agreement for epochs where the rule-based algorithm agreed versus disagreed with the 10-scorer majority-vote reference. The horizontal axis shows the per-epoch agreement ratio (fraction of scorers assigning the modal sleep stage). Disagreements with the reference are proportionally more common where human agreement is lower.

5.4 Discussion

This work proposes a rule-based automatic sleep staging method that operationalizes the AASM scoring manual [8] as executable code, giving decisions that are fully deterministic and explainable [15], [136]. Rather than treating sleep staging solely as a statistical prediction problem, the method codifies the clinical decision process, such that the reasoning can be inspected and rendered as static or interactive natural-language explanations. In addition to the sleep stage classifier, the method thus provides a tool for auditing, education, and explanation in clinical settings where alignment with standardized scoring rules is of importance.

The method clearly does not outperform contemporary ML and DL methods in terms of scoring accuracy [19], [26], [31], [199], [202]. Our results show that although recall for sleep stages N2 and R is strong, the method showed lower agreement with the reference for Wake, N1 and, to a lesser extent, N3. However, it provides an alternative, complementary way to estimate sleep stages. This opens up the possibility of hybrid automatic sleep staging methods [25], [49], where a deterministic and auditable method, grounded in clinical rules, runs in parallel with ML/DL sleep staging methods, as a reference or cross-check.

The reviewer-based discrepancy assessment performed in this work clarifies where the current implementation differed from the majority-vote reference and how those differences arose. Many disagreements between the algorithm and the human consensus scoring occurred in epochs where human inter-scorer agreement was also low, suggesting that disagreements often occur in ambiguous regions of sleep. The assessment also showed that when the reviewer disagreed with the algorithm's sleep stage label, the micro-annotations were judged to be incorrect, pointing to the micro-annotation quality as the main technical limitation of the current implementation.

The micro-annotations were not independently validated against expert event-level annotations. In particular, alpha rhythm and LAMF were operationalized using PSD-based heuristics rather than a separately validated detector. Future work should therefore validate and improve the available micro-annotations, including alpha rhythm, LAMF, REMs, spindles, SWA, chin EMG tone, and missing detectors such as K-complex detection. Wake classification was particularly difficult to operationalize. This may

reflect that some visual cues used by human scorers to recognize Wake epochs are less straightforward to translate into computational rules, especially in noisy epochs with movement artifacts. The scoring manual is written to support expert human judgment but leaves ambiguity for implementation [15]. For example, some scoring criteria depend on judgment of whether the majority of an epoch contains evidence for a sleep stage, without specifying how that evidence should be segmented or aggregated computationally [8]. Future iterations of scoring manuals could benefit from more operationally precise phrasing, fit for algorithmic implementation.

5.5 Conclusion

Taken together, these findings suggest that even if rule-based sleep staging is not yet competitive with DL in terms of agreement with human-scored references, it remains valuable as a transparent complement to black-box methods. Its current limitations appear in the lack of reliable micro-annotation and the ambiguity of a scoring manual written for human experts.

Chapter 6

Error propagation to sleep metrics

Error propagation from sleep stage classification to derived sleep parameters in machine learning on data from wearables

Full reference: E. Hardarson *et al.*, 2023, "Error Propagation from Sleep Stage Classification to Derived Sleep Parameters in Machine Learning on Data from Wearables", *Current Sleep Medicine Reports*, vol. 9, DOI: 10.1007/s40675-023-00253-w

6.1 Introduction

Objective sleep measurements are a critical part of sleep research and diagnosis of sleep disorders [207]. Polysomnography (PSG) is the gold standard of objective sleep measurements, but analysis of PSG requires expert human annotation. In addition to being expensive to analyze, PSG is time-consuming to set up, requires specialized equipment and causes discomfort, making prolonged ambulatory sleep measurements infeasible. To mitigate these limitations, two avenues can be identified in both research and clinical measurements. On the one hand, to ease the data collection for a

broader part of the population, consumer-grade wearables have become a prominent tool [208], and on the other hand, automation of the analysis through the development of machine learning models [9], [26], [27].

Machine learning models, and deep learning, in particular, are commonly used to classify sleep stages (sleep staging). Of the 185 Google Scholar article results where both “sleep" and “deep learning" were present in the title, approximately 40% were about sleep staging (similar findings for PubMed, author search October 20^{th}, 2022). Sleep staging is an important step in the analysis and diagnosis of sleep because it establishes the macrostructure of sleep, and several sleep statistics are derived from the sleep stages [209]. These clinically relevant sleep metrics include, for example, sleep efficiency (SE), wake after sleep onset (WASO) and amount of time spent in rapid eye movement (REM) sleep. Machine learning models typically follow the same protocol as human scorers and designate each 30-second non-overlapping time segment (epoch) one sleep stage. This results in a sequence of sleep stages called a hypnogram which is then compared to a hypnogram where an expert human scorer performed the sleep staging, to evaluate the accuracy of the machine learning model. In this chapter, we refer to this manually scored hypnogram as the “ground-truth hypnogram".

Most published work describing machine learning models for classifying sleep stages only reports the sleep staging error of the model, with the implicit assumption that this would translate to a lower error in derived sleep metrics. However, it is unclear how robust sleep metric estimation is to misclassifying sleep stages and to the best of our knowledge, the impact of sleep staging error on the derived sleep metric error is poorly understood. For example, it is not known how well a machine learning model capable of 92% accurate sleep staging performs when estimating sleep efficiency. This chapter focuses on the propagation of error from sleep stages to sleep metrics.

The mapping from a hypnogram to sleep metrics is highly non-linear and depends on multiple interacting factors, not least the structure of the underlying ground-truth hypnogram. Thus, a formulation of error propagation cannot be derived analytically. Instead, here, we propose a Monte Carlo error propagation simulation framework tailored to the analysis of sleep data. We simulated how a machine learning model would estimate sleep stages for a

dataset of ground-truth hypnograms, given the machine learning model's confusion matrix.

Uncertainty in sleep staging can be attributed to several diverse sources, with recent developments in quantifying and describing this uncertainty [210], [211]. However, the literature is scarce regarding the error in sleep metrics derived from a hypnogram. Although not commonly applied in sleep science, Monte Carlo simulation is a well-established method for evaluating error propagation in other fields, such as physics, biology, chemistry and geology [212]–[214]. In this chapter we present the results of a Monte Carlo simulation study showing the impact of error propagation relative to the error of sleep staging models.

Wearable sleep trackers, such as wrist-worn smartwatches, have been suggested as an alternative or complement to the resource-heavy PSG to measure sleep, e.g. for screening of sleep disorders [207]. The reason is two-fold: i) wearables provide light-weight measures at a relatively low cost compared to the PSG, and; ii) wearables can continuously capture data over a long time, which enables new types of analysis to be conducted, allowing measurements of so-called free-living sleep [215]. However, there is a gap in the literature regarding what type of data they can capture and how to analyze the data.

Most consumer-grade wearables currently in use are wrist worn [216], [217]. The technology is promising but requires validation, access to the algorithms applied to the aggregated data or access to the raw data. Furthermore, wearable sleep trackers are less accurate than PSG at classifying sleep stages [207], [208], [218]–[222]. Hence, understanding how staging error propagates to sleep metrics is especially important when working with data from wearable sleep trackers.

With this background, we opted to limit the scope of this simulation study to examine machine learning models which estimate sleep stages from wrist-worn wearables. This requires finding reported confusion matrices and other details of such machine learning models.

Our main contributions are three-fold:

1. A review of machine learning models using data from wrist-worn sleep trackers, demonstrating a gap in the literature when it comes to reporting error statistics.

2. A novel Monte Carlo simulation framework for estimating

error propagation from sleep staging error to clinical sleep metrics, which demonstrates that sleep staging error is not representative of the error in derived sleep metrics.

3. Suggestions for future work developing machine learning models for estimating clinically relevant sleep metrics from wearable data.

6.2 Background

Previous work has reviewed the use of wearables to detect sleep stages [221], [223], [224] and the use of machine learning models to detect sleep stages [9], [27], [28], [225]. We searched the literature for papers about machine learning models performing sleep staging on data from wrist-worn wearables. Performing a Web of Science search for the query

```
(ALL=(wearable OR wearables OR wrist)) AND
(ALL=(stages OR staging OR stage)) AND
ALL=(classifier OR machine learning OR
deep learning OR probabilistic OR
logistic regression) AND (TI=sleep)
```

resulted in 39 papers, of which 30 were excluded because they did not fit the following eligibility criteria:

- Does the article present a machine learning model for automatic sleep staging?
- Does the machine learning model use data from a wrist-worn wearable device?
- Is the machine learning model's accuracy reported?
- Does the paper describe an original machine learning model? Review articles are excluded.

An identical search on PubMed did not offer new results. The search was performed on August 5^{th}, 2022. Table 6.1 summarizes the findings of the literature search regarding which sensors were used, how many sleep stages were considered, the machine learning model used, the staging accuracy and whether the paper reported the confusion matrix or information about sleep metrics. As can be seen in the table, most studies use photoplethysmography (PPG),

acceleration (ACC) or electrodermal activity (EDA); their accuracy ranges from 59% to 93%; five out of nine papers show the confusion matrix and only two report information about sleep metrics. Eight of the nine papers used 30-second epochs; Gashi et al. classified 600-second segments [226]. Only Wulterkens et al. and Fonseca et al. reported the error in clinical sleep metrics [227], [228].

Epochs can be classified into various numbers of classes. Common designations are 2-class (wake or sleep), 3-class (wake, N1/N2/ N3 or REM sleep), 4-class (wake, N1/N2, N3 or REM sleep) and 5-class (wake, N1, N2, N3 or REM sleep). When 4-class staging is performed, the N1/N2 class is commonly referred to as *light sleep* and the N3 class as *deep sleep*. Of the nine machine learning models, five performed 4-class sleep staging These five techniques form the basis of our simulation study. All five techniques use data from research devices, as opposed to consumer-grade wearables. Comparing the staging performance of machine learning models is problematic since the accuracy depends on multiple factors, such as the dataset used, which sensors were used, how lights out was defined and what validation method was used. For example, the machine learning model described by Anusha et al. has the highest reported 4-class staging accuracy in Table 6.1. However, it was validated against a home sleep apnea testing device with no electroencephalography data recorded. [229].

6.2.1 Sleep metrics

The American Academy of Sleep Medicine (AASM) Scoring Manual Version 2.6 lists the following recommended metrics for polysomnography which must be reported for clinical sleep studies [209]:

- *Total recording time (TRT).* Time in minutes between the first epoch of any sleep and last epoch of any sleep. Since the dataset used in the simulation did not contain information about lights on and lights off, this metric and those derived from it were defined differently from the AASM definition.
- *Total sleep time (TST).* Time in minutes spent in light, deep and REM sleep.
- *REM latency.* Time in minutes between sleep onset to the first epoch of REM sleep.
- *Wake after sleep onset (WASO).* TRT - TST.

Table 6.1: A summary of published papers using machine learning for sleep staging on data from wrist-worn wearables, in order of publication date.

Reference	Sensors	Stages	Accuracy (κ)	ML method	C[1]	y[2]
Gashi et al. 2022 [226]	EDA, ST, ACC	2	92% (0.55)	GB		
Anusha et al. 2022 [229]	EDA, ST	4	93% (-)	RF	x	
Ko et al. 2022 [230]	HR, ACC	2	85% (-)	CK		
Radha et al. 2021 [231]	PPG	4	76.36% (0.65)	LSTM	x	
Wulterkens et al. 2021 [227]	PPG, ACC	4	76.4% (0.62)	LSTM	x	x
Li et al. 2021 [232]	PPG, ACC	4	69% (0.44)	CNN, SVM	x	
Sundararajan et al. 2021 [233]	ACC	2	79%[3] (0.5)	RF	x	
Walch et al. 2019 [234]	ACC, PPG	3	72.3% (0.27)	MLP		
Fonseca et al. 2017 [228]	PPG	4	59% (0.42)	LDC		x[4]

Abbreviations: ACC, acceleration; C, confusion matrix; CK, Cole–Kripke; CNN, convolutional neural network; EDA, electrodermal activity; GB, gradient boosting; HR, heart rate; LDC, linear discriminant classifier; LSTM, long short-term memory; ML, machine learning; MLP, multilayer perceptron; PPG, photoplethysmography; RF, random forest; ST, skin temperature; SVM, support vector machine; y, sleep metric.
[1] Marked "x" if a confusion matrix was reported.
[2] Marked "x" if information about estimated sleep metrics was reported.
[3] Average precision.
[4] Fonseca et al. reported estimated sleep metrics, but only using a 2-class sleep/wake classifier.

- *Sleep efficiency (SE).* Proportion of time spent sleeping: TST/TRT.
- *Time in minutes in each state.* The four included stages are wake, light, deep and REM sleep.
- *Proportion in each state.* Percentage of TST in each stage (time in each stage/TST).

The simulation study in this chapter uses four sleep stages (wake, light, deep and REM), so time in each state and proportion in each sleep state leads to a total of six metrics. In addition to the AASM

recommended metrics listed above, the number of awakenings (NA) is defined as the number of times at least two subsequent wake epochs occur after sleep onset [235]. As a result, we consider in total eleven sleep metrics.

6.3 Methods

Performance of machine learning sleep stage classifiers is most commonly reported in terms of staging accuracy, Cohen's kappa and sleep stage confusion matrix. Sleep metric error is not commonly reported and can not be derived analytically from the staging error. To gain an understanding of how staging error propagates to sleep metrics, we performed a simulation study [236]–[238] where we resampled ground-truth hypnograms and monitored how the derived sleep metrics change. We note that in what follows, we used the error of sleep staging models instead of their accuracy, where error $= 1 -$ accuracy. We refer to this as *staging error*.

The simulation study is based on the Sleep-EDF Database Expanded (Sleep-EDFx) dataset [159], [239], [240], which contains 197 PSGs scored by human experts according to the Rechtschaffen and Kales manual [14]. The data were obtained in two studies of mostly healthy adults. We refer to the original publications for more details on the composition of the subject group [239], [240]. We treated these 197 hypnograms, which we refer to as $x_1, ..., x_{197}$, as ground-truth hypnograms and use them to calculate ground-truth sleep metrics. Using a Monte-Carlo approach, we simulated the operation of a staging classifier with a given confusion matrix C by repeatedly resampling the hypnograms: We pseudo-randomly selected a new sleep stage for each epoch based on confusion matrix C. We then calculated the sleep metrics according to the simulated hypnogram and the residual error, i.e. the difference to the ground truth sleep metric value. This was repeated n_{sim} times for every hypnogram. This gave us an estimated sleep metric error for a given confusion matrix. The process is described in Algorithm 2 and schematically in Fig. 6.1.

Stated mathematically (following the notation in [236]), we considered the calculation of a sleep metric y given a hypnogram x by computational model $\mathcal{M} : \mathcal{X} \to \mathcal{Y}$ where the random variable $X \in \mathcal{X}$ with probability density function $p_X(x)$ represents a hypnogram and $Y = \mathcal{M}(X)$ is a random variable representing the

calculated sleep metric.

Typically, uncertainty propagation is estimated by studying statistics such as the variance of the output variable, $Y \in \mathcal{Y}$. Since we treated the human expert scorer's classification as the ground truth, we instead considered the residual error, $e_y = y - \hat{y}$, that is the difference between the ground truth metric value and the metric value resulting from the simulation (sim), given a confusion matrix, C. The Monte Carlo estimation of sleep metric error e_y is given by

$$\hat{e}_y = \frac{\sum_i^{n_{\text{sim}}} (y - \hat{y}_i)}{n_{\text{sim}}}$$

and the estimated root mean squared error (RMSE) and mean absolute error (MAE) over the dataset are given by

$$\text{RMSE}_y = \frac{\sum_i^{n_{\text{sim}}} \sqrt{\frac{\sum_j^N (y - \hat{y}_j)^2}{N}}}{n_{\text{sim}}}$$

and

$$\text{MAE}_y = \frac{\sum_i^{n_{\text{sim}}} \frac{\sum_j^N |(y - \hat{y}_j)|}{N}}{n_{\text{sim}}},$$

respectively, where N is the number of hypnograms in the dataset.

6.3.1 Generation of Confusion Matrices

The Monte Carlo method described in Algorithm 2 simulates the resampling of a dataset of hypnograms as though they were estimated with confusion matrix C. We had a few examples of confusion matrices from the literature [227], [229], [231], [232], but in order to properly visualize the relationship between staging error and sleep metric error, we also generated a multitude of hypothetical confusion matrices.

For a given target four-class staging error, $\tilde{e}_{\text{class}} \pm \epsilon$, where ϵ is a small value, we generated a hypothetical confusion matrix, $C_{\tilde{e}}$ as follows: $C_{\tilde{e}}$ is initialized as a 4x4 zero matrix. First, we assign the four elements on the diagonal of $C_{\tilde{e}}$ by sampling from a Beta distribution with a maximum at the target staging accuracy:

$$C_{\tilde{e}}(i,i) = M(i) \cdot k, \qquad k \sim \mathcal{B}(\alpha, \beta)$$

where $M(i)$ is the expected proportion of epochs of sleep stage $i \in \{1, 2, 3, 4\}$. We set

$$\alpha = \left(\frac{1-\mu}{\sigma^2} - \frac{1}{\mu} \right) \cdot \mu^2$$

and

$$\beta = \alpha \cdot \left(\frac{1}{\mu} - 1 \right)$$

where μ is the target staging error and $\sigma^2 = 0.01$. The remaining cells of $C_{\tilde{e}(j)}$ were sampled uniformly from the range $[0; M(j) - \sum C_j]$ in random order. When there was only one cell left in row j it was assigned the value

$$M(j) \cdot \left(1 - \sum C_j \right)$$

This allowed us to visualize the relationship between staging error and sleep metric error by generating thousands of hypothetical confusion matrices and applying Algorithm 2 to each one.

Table 6.2: Example of a generated confusion matrix of 1,000 epochs. This confusion matrix corresponds to a classifier with an overall accuracy of 80% but only 42% REM sleep sensitivity.

		Predicted			
		Wake	Light	Deep	REM
True	Wake	69	15	4	12
	Light	34	532	11	23
	Deep	11	4	135	0
	REM	6	2	79	63

Tables 6.2 and 6.3 show two examples out of 20,000 generated confusion matrices, where $\mu = 0.2$ was used. These are examples of confusion matrices where sensitivity to specific sleep stages are low while overall accuracy is 80%. For the 20,000 generated confusion matrices the overall accuracy was $(80 \pm 6)\%$ while the sensitivity to all individual sleep stages was $(80 \pm 10)\%$. Table 6.4 shows the combined distribution of all 20,000 confusion matrices.

Table 6.3: Example of a generated confusion matrix of 1,000 epochs. This confusion matrix corresponds to a classifier with an overall accuracy of 80% but only 71% light sleep sensitivity.

		Predicted			
		Wake	Light	Deep	REM
True	Wake	89	1	6	4
	Light	111	424	30	35
	Deep	2	1	147	0
	REM	3	4	5	138

Table 6.4: The mean and standard deviation ($\bar{\mu}\pm$STD) of individual cells in 20,000 generated confusion matrices with 1,000 epochs each.

		Predicted			
		Wake	Light	Deep	REM
True	Wake	80 ± 10	7 ± 7	7 ± 7	7 ± 7
	Light	40 ± 42	480 ± 59	40 ± 42	40 ± 42
	Deep	10 ± 11	10 ± 11	119 ± 15	10 ± 11
	REM	10 ± 11	10 ± 11	10 ± 11	119 ± 15

Limitations

The simulation approach used has some shortcomings, which limit the scope of scenarios to which it is applicable. First of all, the resampling step in Algorithm 2 re-assigned each epoch a sleep stage independently of other epochs, so the method assumed that the staging error was random in time. This assumption holds for machine learning models where an epoch is scored without taking into account the epoch's temporal context, i.e., when $p(\hat{x}(k) \mid x(k))$ is independent of $x \setminus x(k)$. This is not the case for all of the machine learning models in Table 6.1. Secondly, the results of the simulation on the Sleep-EDFx dataset cannot be assumed to apply to other datasets unless the structure of participants' sleep is statistically similar to the Sleep-EDFx dataset. Finally, it should be acknowledged that all Monte Carlo methods are based on pseudo-random numbers, and uncertainty in the results is a

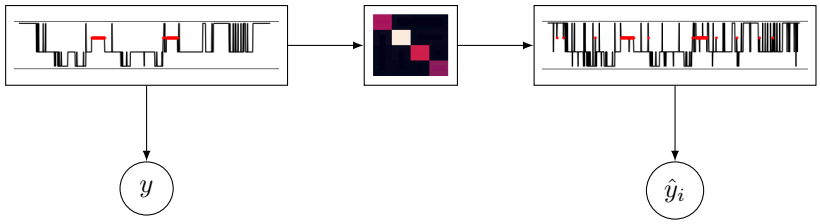


Figure 6.1: A schematic of iteration i of the simulation process, which generates an estimated sleep metric error, e_y. A manually scored hypnogram (ground truth) hypnogram, x, is selected from $\{x_1, ..., x_N\}$, a set of experimentally observed hypnograms. The resampled hypnogram, $\hat{x}_i$ is created by randomly reassigning each 30-second epoch in x a new sleep stage by making a random selection, weighted by row number s of confusion matrix, C, where s corresponds to the index of the true sleep stage. Sleep metrics y and $\hat{y}_i$ are calculated from x and $\hat{x}_i$, respectively and their residual, $e_y = y - \hat{y}_i$ is the sleep metric error.

consequence not only of false assumptions but also of the method's inherent randomness [238].

6.3.2 Baseline models

To put the simulated sleep metric error into perspective, we need a naïve baseline model to compare against. To that end, we split the Sleep-EDFx dataset into a training and testing dataset with an approximately 50/50 split, ensuring that no individual participant's data was found in both sets. Using the training set, we calculated the average of sleep metric y across every hypnogram,

$$\bar{y} = \frac{\sum_j^{N_{\text{train}}} y(x_j)}{N_{\text{train}}}$$

This was our baseline (BL) model for sleep metric y, i.e. $\hat{y}^{\text{BL}} = \bar{y}$. The baseline sleep metric model error was found by applying the baseline sleep metric model to the test set:

$$\text{MAE}_y^{\text{BL}} = \frac{\sum_j^{N_{\text{test}}} \mid y_j - \bar{y} \mid}{N_{\text{test}}}$$

Algorithm 2 Monte Carlo simulated sleep metric error for sleep metric y.

Require: Set of N hypnograms $\{x_1, ...x_N\}$; confusion matrix, C; number of simulation repetitions, n_{sim}; formula for calculating sleep metric, $f(x)$.

1: **procedure** SIMULATEMETRICERROR($\{x_1, ...x_N\}, C$) **for** $i \in [1, n_{sim}]$ **do**
    end
    $j \in [1, N]$
2: Select the j^{th} hypnogram, x_j
3: Calculate the ground-truth metric value, $y_j = f(x_j)$.
4: $M \leftarrow$ number of epochs in x_j **for** $k \in [1, M]$ **do**
5:
    end
    Set weights $w \leftarrow C_{x_j(k)}$, the $x_j(k)^{\text{th}}$ row of C.
6: Generate $\hat{x}_j(k)$ by randomly selecting one of the sleep stages according to weights, w.
7:
8: Calculate estimated sleep metric $\hat{y}_j = f(\hat{x}_j)$
9: $e_{i,j} = y_j - \hat{y}_j$
10:
11:
12: **return** $e_{i,j}$ *or statistics of* e *such as the mean of estimators* e_i *across the* n_{sim} *simulations,* $\hat{e}$.
13: **end procedure**

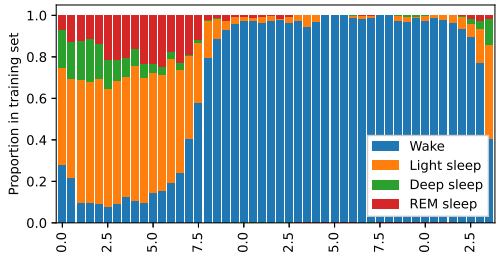


Figure 6.2: Proportion of time spent in each sleep stage, every half hour from midnight, for the subjects in the training set. This information was used in a baseline staging model, which randomly assigned a sleep stage given the time of day to the subjects in the testing set. Abbreviations: REM: Rapid eye movement.

and

$$\mathrm{RMSE}_y^{\mathrm{BL}} = \sqrt{\frac{\sum_j^{N_{\mathrm{test}}} (y_j - \bar{y})^2}{N_{\mathrm{test}}}}$$

We also needed to decide on a range of staging errors for generating confusion matrices. To establish the highest reasonable staging error, we again created a naïve baseline model, this time for sleep staging. Using the test set, we split the dataset into half-hour time slots. We then counted how many epochs were assigned to each sleep stage per half-hour time slot. Fig. 6.2 shows the result. The baseline staging model randomly assigned a sleep stage given the time of day, weighted by the proportional split of sleep stages in the corresponding half-hour time bin.

6.4 Results and discussion

Algorithm 2 produced resampled hypnograms for a given staging error. Fig. 6.3 shows examples of a hypnogram with 30-second epochs resampled by Algorithm 2, using confusion matrices reported in three of the papers in Table 6.1. We see that for a higher staging error, the hypnogram became more chaotic and less representative of actual sleep. However, the same may not necessarily be said about the derived sleep metric error as reported in the literature. Of the machine learning models in Table 6.1,

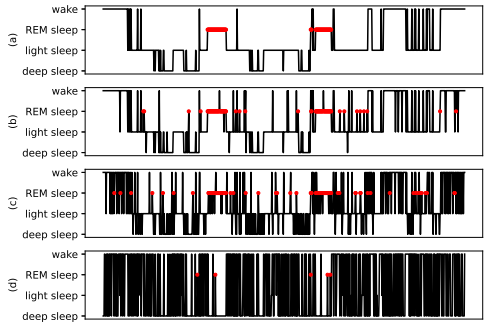


Figure 6.3: The effect of resampling a hypnogram. a) An original manually scored hypnogram and the same hypnogram, resampled using the confusion matrix from b) Anusha et al. 2022 (93% staging accuracy), c) Wulterkens et al. (76% staging accuracy) and d) Li et al. 2019 (69% staging accuracy). Abbreviations: REM: Rapid eye movement.

only Wulterkens et al. reported estimated sleep metric statistics for four-class staging [227].

Fig. 6.4 shows the distribution of residual sleep metric error over the dataset using the same three confusion matrices. In blue, is the simulation of the staging in the paper by Wulterkens et al. (76% staging accuracy), and the dotted lines show the upper and lower limits of the error reported in the paper. When comparing the simulated error distribution to the error range reported by the authors (dotted lines in Fig. 6.4) we found that the simulation agreed reasonably well with their empirical results.

The simulated resampling of a low-error staging machine learning model resulted in a low-variance, low-bias sleep metric estimation and as the staging error increased, so did the sleep metric bias and variance (Fig. 6.4). Comparing the reported and simulated RMSE of the machine learning model from Wulterkens et al. listed in Table 6.5, we again found a reasonably good agreement.

6.4.1 Baseline values

We applied the baseline models for sleep staging and sleep metrics to the test set. This gave us baseline error values to compare the estimated sleep metric error with. To be considered relevant, the models had to have an error value that was below this threshold.

Table 6.5: RMSE mean and standard deviation for sleep metric error residual from ground truth, as reported by Wulterkens et al. and the estimate from Monte Carlo simulation. The rows where reported values are missing correspond to metrics not reported by Wulterkens et al.

	Error mean ± SD		RMSE	
Metric	Reported	Simulated	Reported	Simulated
WASO [min]	9.8 ± 39.5	27.3 ± 46.8	40.6	54.2
TST [min]	-4.0 ± 34.1	22.2 ± 42.3	34.3	47.8
SE [%]	-0.7 ± 6.6	2.5 ± 4.6	6.7	5.3
Time in light sleep [min]	1.2 ± 50.0	6.4 ± 40.2	49.9	40.7
Proportion light sleep [%]	0.6 ± 10.7	-2.2 ± 5.9	10.7	6.3
Time in deep sleep [min]	1.2 ± 40.3	9.4 ± 17.9	40.2	20.2
Proportion deep sleep [%]	-0.6 ± 9.6	1.7 ± 4.3	9.6	4.7
Time in REM sleep [min]	-6.4 ± 22.9	6.4 ± 11.1	23.8	12.8
Proportion REM sleep [%]	-1.2 ± 5.8	0.6 ± 2.7	5.9	2.7
Number of awakenings		42.5 ± 47.1		63.4
REM latency [min]		-126.0 ± 117.2		172.1

Note: The uncertainty value in the simulated error column denotes the estimated error mean and estimated standard deviation, $\hat{\mu}(e)$ and $\hat{\sigma}^2(e)$, not the Monte Carlo standard error. Abbreviations: REM: Rapid eye movement, RMSE: Root mean squared error, SD: Standard deviation, SE: Sleep efficiency, TST: Total sleep time, WASO: Wake after sleep onset.

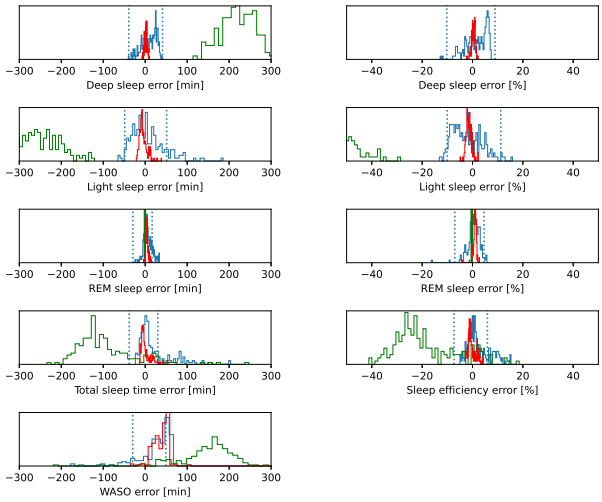


Figure 6.4: Histograms showing the distribution of simulated sleep metric error across the dataset for one iteration of the Monte Carlo method ($n_{\text{sim}} = 1$). The blue shows a simulation of the staging in the paper by Wulterkens et al. [227] (76% staging accuracy), and the dotted lines show the upper and lower limits of the error reported in the paper. The histogram in red shows a simulated error distribution for the staging in the paper by Anusha et al. [229] (93% staging accuracy), and the green shows a simulation of the machine learning model by Li et al. [241] (69% staging accuracy). Abbreviations: REM: Rapid eye movement, WASO: Wake after sleep onset.

The resulting baseline values can be seen in Table 6.6. This set the lower baseline for 4-class sleep staging accuracy at 46% and a Cohen's kappa of 0.14. Table 6.6 also includes accuracy values for 2-, 3- and 5-class sleep staging baseline models for comparison. Furthermore, Table 6.6 lists baseline error values for sleep metrics in terms of MAE and RMSE. Not shown in Table 6.6 is the error in sleep metrics obtained by using hypnograms estimated using the baseline staging model. This consistently resulted in higher sleep metric errors, than when using the average metric values from the training set.

6.4.2 Clinical relevance of staging error

Fig. 6.5 shows the simulated sleep metric RMSE over the whole dataset as a function of the four-class staging error for the twenty thousand generated confusion matrices, as well as confusion matrices reported in the literature (circles) [227], [229], [231], [241]. Also plotted are the staging error and sleep metric RMSE reported in Wulterkens et al. (triangle) [227], again showing reasonable agreement between simulated and empirical results.[1]

Looking at Fig. 6.5, it is apparent that the relationship between staging error and sleep metric error is non-linear. In general, the relationship seems to be logarithmic, although interestingly, error in the number of awakenings (NA error) does not strictly increase with staging error. This is an artefact of how NA is defined. A very inaccurate staging model will produce a hypnogram similar to the labeled (d) Fig. 6.3, with very few contiguous wake epochs. Since close to zero contiguous wake epochs are found, the RMSE will be approximately equal to the true NA, on average around 13. This behaviour can be seen on the right side of the plot in Fig. 6.5. A more accurate staging model will correctly label more wake epochs, similar to the graph labeled (b) in Fig. 6.3. For prolonged wake bouts, made up of many epochs, some will likely be misclassified as sleep, splitting the wake bout in two. Thus, every misclassification of a wake epoch effectively increases NA by one, resulting in higher NA error, as shown by the peak in

[1] It should be noted that the relationship between kappa and staging error is approximately linear within the domain of four-class sleep staging on a homogeneous population. Thus, the relationship between sleep metric error and Cohen's kappa of the staging machine learning model is similar to that seen in Fig. 6.5.

NA error for staging error between 0.05 and 0.35. In general, some metrics' error grows gradually as staging error increases, but others (particularly WASO, REM latency and NA) seem to be much more sensitive to minor staging errors.

The dotted lines in Fig. 6.5 show the RMSE of the baseline sleep metric estimators. While some sleep metrics (such as WASO and NA) seem to be estimated below baseline error irrespective of staging accuracy, others (such as deep sleep and REM duration) are estimated with an error surpassing baseline error, even when staging error is low.

The simulation shows that a highly accurate staging model (one far surpassing the wearable staging accuracy of today's state of art) could conceivably perform worse than the baseline model on the estimation of derived sleep metrics. Under the assumptions made here, models capable of 90% sleep staging accuracy were, on average, estimated to perform below a baseline model in two out of eleven sleep metrics, essentially making the estimations worse than a random guess. For comparison, the current state-of-the-art in machine learning 5-class sleep staging on PSG data is around 85% [26], [210].

In recent years, a large body of literature has focused on developing machine learning models to automate the analysis of sleep measurements in order to ease the labour-intensive analysis and allow the use of wearables and other screening devices [10], [160], [210]. However, these models are seldom used in clinical practice [26], [68].

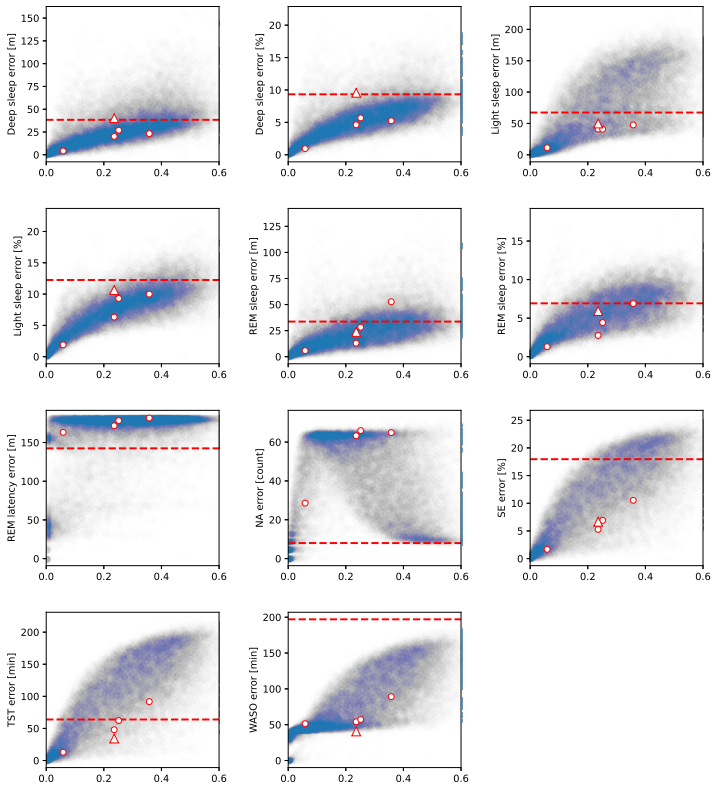


Figure 6.5: The RMSE of different simulated sleep metrics as a function of the staging error of the confusion matrix used. The blue semi-transparent dots show simulations of 20,000 generated confusion matrices, and the red circles show simulations of confusion matrices reported in the literature [227], [229], [231], [241]. The red triangles show the RMSE of reported sleep metrics for a given staging accuracy from Wulterkens et al. [227]. The dotted line shows the RMSE of the baseline sleep metric estimation model. All data points which lie above the dotted line can be interpreted as less accurate than an educated guess. Abbreviations: NA: Number of awakening, REM: Rapid eye movement, RMSE: Root mean squared error, SE: Sleep efficiency, TST: Total sleep time, WASO: Wake after sleep onset.

Table 6.6: Performance of baseline models for staging and sleep metric estimation. The staging model randomly assigns a sleep stage to an epoch according to the time of day. The metric estimation model always assigns the average metric value of the training set.

Metric	Baseline model performance
	Accuracy (Cohen's kappa)
Wake/sleep staging	74% (0.25)
Wake/NREM/REM staging	54% (0.15)
Wake/light/deep/REM staging	46% (0.14)
Wake/N1/N2/N3/REM staging	35% (0.10)
	MAE (RMSE)
WASO [min]	135.1 (197.0)
TST [min]	47.2 (64.0)
Number of awakenings [count]	6.5 (8.0)
SE [%]	14.1 (18.0)
REM Latency [min]	90.0 (142.5)
Time in light sleep [min]	52.7 (67.3)
Proportion light sleep [%]	9.9 (12.3)
Time in deep sleep [min]	32.2 (38.4)
Proportion deep sleep [%]	7.9 (9.3)
Time in REM sleep [min]	25.5 (33.7)
Proportion REM sleep [%]	5.2 (6.9)

Abbreviations: MAE: Mean absolute error, NREM: Non-rapid eye movement, REM: Rapid eye movement, RMSE: Root mean squared error, SE: Sleep efficiency, TST: Total sleep time, WASO: Wake after sleep onset.

6.5 Conclusions

We have presented a simulation framework for exploring how sleep staging error propagates to the calculations of sleep metrics. The framework was evaluated using an open-source dataset of sleep studies and four-stage classification metrics from studies using wearable sleep trackers. However, the framework can be expanded to apply to other scenarios, such as five-class sleep staging and machine learning on PSG data. Future work could use this framework to investigate sleep metric error propagation in other approaches, such as EEG or ECG, where sleep metric error is more widely reported.

Clinically, sleep staging accuracy is an irrelevant metric by itself, yet it is the most popular task for training machine learning models in the literature. The results of the simulations performed here may indicate that machine learning models should not use staging accuracy as a target variable because of how the error propagates from sleep stages to sleep metrics. If the underlying purpose of estimating sleep stages is to calculate sleep metrics, it may prove more effective and robust to train regression models to predict sleep metrics directly and to report sleep metrics and their errors when training sleep staging models. This could lead to more accurate results, which is important for clinical practice.

Treating sleep metric estimation as a regression task poses some new challenges. Because sleep metrics are typically properties of a whole night of sleep, in contrast to sleep stages, which are properties of individual epochs, the training data needed for sleep metric regression may be much larger than for training sleep staging models. Furthermore, not all clinically important information is contained in sleep metrics. For example, following a sleep study, a clinician would typically review the hypnogram in order to get an overview of the sleep architecture. However, epoch-by-epoch sleep stage classification may not be the best way to generate the hypnogram and the staging error may not be indicative of the hypnogram error, as shown in Fig. 6.3.

In conclusion, this simulation study explores nuances around the topic of error in sleep metrics derived from an estimated hypnogram. Future work performing automated analysis of sleep data should not focus solely on attaining higher sleep staging accuracy. Error statistics of sleep metrics and other clinically relevant quantities should be reported. Machine learning experts

and sleep scientists should have a dialogue about alternative goals for the future.

Chapter 7

Inter-scorer reliability in sleep architecture metrics

Full reference: (Manuscript in preparation.) E. Hardarson *et al.*, 2026, "Inter-scorer reliability in sleep architecture metrics"

7.1 Introduction

Sleep architecture metrics such as total sleep time (TST) and time spent in non-rapid eye movement (NREM) sleep stages are widely used to describe sleep recorded by polysomnography (PSG). These quantities are routinely used in sleep research, clinical reporting, and the interpretation of sleep continuity and structure [7], [209].

Under current scoring practice, these metrics are not measured directly from physiological signals. Instead, PSG recordings are segmented into 30-second epochs, each epoch is assigned a discrete sleep stage according to standardized scoring rules, and summary sleep metrics are then computed from the resulting hypnogram [14], [209]. This workflow has become the standard operationalization of sleep architecture in both manual and automated sleep analysis.

However, manual sleep staging is known to be imperfect. Epoch-by-epoch inter-scorer disagreement of around 20% has been documented for decades and is generally treated as an expected feature

of manual scoring rather than an exception [12], [26], [194].

The clinical and scientific quantities of interest are often not the epoch labels themselves, but rather the derived sleep architecture metrics. A key unresolved question is, therefore, how disagreement at the sleep stage level propagates to disagreement in the downstream quantities, i.e. sleep architecture metrics, that are actually interpreted and reported. Small differences in epoch-by-epoch staging may accumulate into meaningful differences in derived metrics such as TST, rapid eye movement (REM) duration, or sleep stage 3 (N3) duration. Conversely, hypnograms that appear broadly similar may still yield materially different summary metrics. Our previous simulation work using wearable data suggested that this propagation can be nonlinear and strongly metric-dependent (see Chapter 6) but empirical characterization in PSG studies using multiple independent human scorers has been limited [18].

This issue is important for at least two reasons. First, if derived sleep metrics themselves have substantial inter-scorer variability, then they should not be treated as fixed ground-truth quantities when interpreting individual PSG reports or comparing cohorts. Second, the evaluation of alternative methods of measuring sleep architecture, such as wearable sleep trackers and automatic scoring algorithms, critically depends on what is considered the reference target. If the same recording yields different plausible metric values across expert scorers, then agreement with a single scorer may be an incomplete or misleading measure of performance [15].

A further complication is that there is more than one reasonable way to construct a consensus value for a sleep metric when multiple scorers are available. One may first derive the metric separately from each scorer's hypnogram and then average the resulting values. Alternatively, one may construct a consensus hypnogram, for example, by epoch-by-epoch majority vote, and derive the sleep metric from that hypnogram. These two procedures are mathematically distinct and may produce different result. Despite this, the implications of consensus construction for sleep architecture metrics have received little explicit attention.

In this study, we examine inter-scorer reliability at the level of commonly used sleep architecture metrics using the Sleep Revolution Multicenter Scoring (SRMS) dataset, in which 50 full-length PSG recordings were independently scored by 10 experienced scorers from multiple sleep centers according to a common scoring protocol [17]. For each scorer and recording, we derived sleep-

stage-duration metrics (N1, N2, N3 and REM duration; minutes), sleep-stage-percentage metrics (N1%, N2%, N3% and REM%), TST, sleep efficiency (SE), sleep onset latency (SOL), REM onset latency (ROL), and wake after sleep onset (WASO). Specifically, we ask three related questions: (i) how variable are these metrics across expert scorers; (ii) how does disagreement in epoch-by-epoch sleep staging relate to disagreement in derived sleep metrics; and (iii) given inter-scorer variability, how should reference values for sleep metrics be defined for method evaluation?

By addressing these questions, we aim to clarify the reliability of standard hypnogram-derived sleep architecture metrics and to make explicit the uncertainty inherent in the quantities that are often treated as reference targets in sleep research and sleep medicine.

7.2 Methods

7.2.1 Dataset and scoring protocol

We analyzed the SRMS dataset, a multicenter PSG scoring dataset designed to assess inter-scorer agreement in sleep staging [17]. The dataset comprises 50 overnight full-length PSG recordings collected between February and June 2021 using Nox A1 devices (Nox Medical, Reykjavik, Iceland), a type II portable PSG system. The study was approved by National Bioethics Committee of Iceland (21–070). All study subjects gave a written informed consent to participate in the study. Each recording was independently scored by 10 experienced scorers from seven sleep centers using a common software environment and a shared scoring protocol based on the American Academy of Sleep Medicine (AASM) rules (version 2.6) [17], [209].

The study population was stratified to include recordings from individuals at risk of obstructive sleep apnea, restless legs syndrome, insomnia, and individuals without identified risk factors, as described previously for the SRMS dataset [17]. Table 7.1 shows a summary of demographic and clinical characteristics of the SRMS cohort.

Table 7.1: Cohort characteristics of the SRMS dataset.

Characteristic	Value
Total number of subjects	50
Male/female subjects	29/21
Subjects with OSA risk	21
Subjects with RLS risk	29
Subjects with insomnia risk[a]	17
No-risk subjects[a]	9
Age, years	42.9 ± 13.7
BMI, kg/m^2	27.3 ± 5.8
Total sleep time, minutes	426.3 ± 61.8
Apnea–hypopnea index, events/h	15.2 ± 15.6

[a] Insomnia risk data were missing from three subjects. Subjects were defined as no-risk if they had no obstructive sleep apnea, restless legs syndrome, or insomnia risk.

Continuous variables are reported as mean ± standard deviation.

Abbreviations: SRMS, Sleep Revolution Multicenter Scoring; OSA, obstructive sleep apnea; RLS, restless legs syndrome; BMI, body mass index.

7.2.2 Definitions of sleep architecture metrics

The derived sleep metrics analyzed in this study follow standard AASM sleep-architecture terminology and were computed from each scorer's hypnogram [209]. We considered total TST; minutes spent in N1, N2, N3, and REM sleep; pooled N1/N2 duration (minutes); sleep stage percentages relative to total sleep time (N1%, N2%, N1/N2%, N3%, and REM%); SE; SOL; ROL; and WASO.

Formally, for each scorer A, let

$$H^{(A)} = \left(H_1^{(A)}, \ldots, H_T^{(A)}\right), \qquad H_t^{(A)} \in \mathcal{S},$$

denote the scored hypnogram for a recording with T epochs, where each epoch has duration $\Delta t = 30$ s, and where $\mathcal{S} = \{\mathrm{W}, \mathrm{N1}, \mathrm{N2}, \mathrm{N3}, \mathrm{REM}\}$ is the set of sleep stages.

Let

$$\mathcal{S}_{\text{sleep}} = \{N1, N2, N3, REM\}$$

denote the set of sleep stages. For each sleep metric π, let $\Psi_\pi(\cdot)$ denote the mapping from a scored hypnogram to the corresponding metric value. For any sleep stage $s \in \mathcal{S}_{\text{sleep}}$, sleep stage duration in minutes is defined as

$$\Psi_s(H^{(A)}) = \Delta t \sum_{t=1}^{T} \mathbf{1}\{H_t^{(A)} = s\}.$$

TST is defined as the total time scored as sleep:

$$\Psi_{\text{TST}}(H^{(A)}) = \Delta t \sum_{t=1}^{T} \mathbf{1}\{H_t^{(A)} \in \mathcal{S}_{\text{sleep}}\}.$$

The pooled N1/N2 duration is defined as

$$\Psi_{\text{N1/N2}}(H^{(A)}) = \Psi_{N1}(H^{(A)}) + \Psi_{N2}(H^{(A)}).$$

Sleep stage percentages are computed relative to TST:

$$\Psi_{s,\%}(H^{(A)}) = 100 \times \frac{\Psi_s(H^{(A)})}{\Psi_{\text{TST}}(H^{(A)})}, \qquad s \in \mathcal{S}_{\text{sleep}}.$$

No explicit lights-off or lights-on markers were available in the dataset. Therefore, time in bed (TIB) is defined as the full recording duration:

$$\Psi_{\text{TIB}}(H^{(A)}) = T\Delta t.$$

SE is then computed as

$$\Psi_{\text{SE}}(H^{(A)}) = 100 \times \frac{\Psi_{\text{TST}}(H^{(A)})}{\Psi_{\text{TIB}}(H^{(A)})}.$$

Sleep onset is defined as the first epoch scored as sleep:

$$\tau_{\text{sleep}}^{(A)} = \min\{t : H_t^{(A)} \in \mathcal{S}_{\text{sleep}}\}.$$

SOL was defined from the recording start to the first sleep epoch:

$$\Psi_{\text{SOL}}(H^{(A)}) = \Delta t\, \tau_{\text{sleep}}^{(A)}.$$

The first REM onset is defined as

$$\tau^{(A)}_{\text{REM}} = \min\{t \geq \tau^{(A)}_{\text{sleep}} : H^{(A)}_t = REM\}.$$

ROL was defined from sleep onset to the first REM epoch:

$$\Psi_{\text{ROL}}(H^{(A)}) = \Delta t\,(\tau^{(A)}_{\text{REM}} - \tau^{(A)}_{\text{sleep}}).$$

WASO was computed from sleep onset to recording end:

$$\Psi_{\text{WASO}}(H^{(A)}) = \Delta t \sum_{t=\tau^{(A)}_{\text{sleep}}}^{T} \mathbf{1}\{H^{(A)}_t = W\}.$$

All duration-based metrics were converted from number of epochs to minutes.

7.2.3 Staging agreement

Epoch-by-epoch inter-scorer agreement was evaluated for three staging formulations.

First, agreement was computed for the original 5-class sleep staging using the manual labels.

$$S = \{W, N1, N2, N3, REM\}.$$

Second, agreement was computed for a pooled 4-class sleep staging in which N1 and N2 were combined into a single category:

$$S_4 = \{W, N1/N2, N3, REM\}.$$

Third, we evaluated a set of binary one-vs-rest sleep staging, defined separately for Wake, N1, N2, N3, and REM, by relabeling each epoch as belonging either to the target sleep stage or to any other sleep stage. For each sleep staging formulation, inter-scorer agreement was summarized using Krippendorff's α [242]. This coefficient is defined as

$$\alpha = 1 - \frac{D_o}{D_e},$$

where D_o is the observed disagreement and D_e is the disagreement expected by chance. An α value of 1 indicates perfect agreement, $\alpha = 0$ indicates agreement no better than chance, and negative values indicate systematic disagreement.

For epoch-by-epoch sleep staging, Krippendorff's α was computed using the nominal difference function[242]. For complete epoch-by-epoch nominal data with a fixed number of scorers per epoch, nominal Krippendorff's α and Fleiss's κ [243] are nearly identical, differing only by a negligible finite-sample correction in the expected-disagreement term. We also confirmed this empirically: the two coefficients agreed to the reported precision for all epoch-by-epoch staging formulations. Only Krippendorff's α is reported, but the results are equivalent for Fleiss's κ.

7.2.4 Agreement of sleep architecture metrics

For each derived sleep metric, the unit of analysis was a single recording, with 10 scorer-specific metric values available for each recording. Inter-scorer agreement for continuous sleep metrics was summarized using intraclass correlation coefficients (ICCs) [244].

We report ICC(2,1), corresponding to a two-way random-effects, absolute-agreement, single-rater formulation (ICC(A,1)); ICC(2,10), corresponding to a two-way random-effects, absolute-agreement, average-rater formulation (ICC(A,10)); and ICC(3,1), corresponding to a two-way mixed-effects, consistency, single-rater formulation (ICC(C,1)) [245], [246].

These complementary coefficients were included to characterize different reliability perspectives. ICC(2,1) quantifies the reliability of a single scorer under an absolute-agreement criterion, ICC(2,10) quantifies the reliability of the mean rating across the 10 scorers, and ICC(3,1) quantifies single-scorer consistency when scorer-specific offsets are not treated as disagreement in the same way as under absolute-agreement formulations[245].

As a complementary non-ICC reliability measure, Krippendorff's α was also computed for each derived sleep metric using the interval difference function [242]. Whereas ICCs describe how much of the variation in continuous metric values is attributable to differences between recordings rather than scorer disagreement, Krippendorff's α compares observed scorer disagreement with the disagreement expected by chance. A further advantage is that the same α framework can be used for both categorical epoch labels and continuous sleep metrics by changing the difference function.

7.2.5 Leave-one-scorer-out error

To quantify how far an individual scorer's derived sleep metrics deviated from the remaining scorer panel, we performed a leave-one-out analysis at the metric level. For each recording i, scorer A, and sleep metric π, let

$$Y_{\pi,i}^{(A)} = \Psi_\pi\left(H_i^{(A)}\right)$$

denote the value of metric π derived from scorer A's hypnogram. The corresponding leave-one-out panel reference was defined as the mean of the same metric across the other nine scorers:

$$\bar{Y}_{\pi,i}^{(-A)} = \frac{1}{K-1} \sum_{\substack{B=1 \\ B \neq A}}^{K} Y_{\pi,i}^{(B)},$$

where $K = 10$ is the number of scorers.

The absolute error and squared error of scorer A relative to the leave-one-out panel reference for recording i were defined as

$$e_{\pi,i}^{(A)} = \left| Y_{\pi,i}^{(A)} - \bar{Y}_{\pi,i}^{(-A)} \right|,$$

and

$$q_{\pi,i}^{(A)} = \left(Y_{\pi,i}^{(A)} - \bar{Y}_{\pi,i}^{(-A)} \right)^2.$$

For each scorer A, scorer-specific mean absolute error (MAE) and root mean squared error (RMSE) were computed across the $N = 50$ recordings:

$$\mathrm{MAE}_\pi^{(A)} = \frac{1}{N} \sum_{i=1}^{N} e_{\pi,i}^{(A)},$$

$$\mathrm{RMSE}_\pi^{(A)} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} q_{\pi,i}^{(A)}}.$$

For each metric, the averages of scorer-specific MAE and RMSE values across the 10 scorers are given by

$$\text{Avg. MAE}_\pi = \frac{1}{K} \sum_{A=1}^{K} \mathrm{MAE}_\pi^{(A)}, \qquad \text{Avg. RMSE}_\pi = \frac{1}{K} \sum_{A=1}^{K} \mathrm{RMSE}_\pi^{(A)}.$$

The reported MAE and RMSE ranges denote the minimum and maximum scorer-specific mean errors across the 10 scorers.

7.2.6 Naive constant predictor comparison

As a simple benchmark for the leave-one-out metric-error analysis, we compared scorer-versus-panel errors with the error of a naive constant predictor. For each sleep metric π, the naive predictor outputs a fixed constant value c_π.:

$$\hat{Y}_{\pi,i}^{\text{naive}} = c_\pi \qquad \text{for all recordings } i.$$

These constants were not estimated from the SRMS dataset. Instead, they were based on global mean baselines from Boulos *et al.* [247], representing average sleep architecture in a broader population. The following constants were used: TST = 394.6 min, N1 duration = 31.2 min, N2 duration = 202.8 min, N1/N2 duration = 234.0 min, N3 duration = 80.5 min, REM duration = 75.0 min, N1% = 7.9%, N2% = 51.4%, N1/N2% = 59.3%, N3% = 20.4%, REM% = 19.0%, WASO = 48.2 min, and SE = 85.7%. Boulos et al. did not publish minute-based sleep stage durations, so these were derived from pooled mean TST and pooled stage percentages, and the pooled N1/N2 duration and N1/N2% constants were obtained by summing the corresponding N1 and N2 averages. No comparison to naive baseline was made for SOL or ROL.

For scorer A, the absolute error and squared error of the naive predictor for recording i are defined as

$$e_{\pi,i}^{(A),\text{naive}} = \left| Y_{\pi,i}^{(A)} - c_\pi \right|,$$

and

$$q_{\pi,i}^{(A),\text{naive}} = \left(Y_{\pi,i}^{(A)} - c_\pi \right)^2,$$

where $Y_{\pi,i}^{(A)}$ denotes the value of metric π derived from scorer A's hypnogram for recording i.

Naive-predictor mean absolute error (MAE) and root mean squared error (RMSE) across the full set of scorer-recording pairs are given by

$$\text{MAE}_\pi^{\text{naive}} = \frac{1}{NK} \sum_{i=1}^{N} \sum_{A=1}^{K} e_{\pi,i}^{(A),\text{naive}},$$

$$\text{RMSE}_\pi^{\text{naive}} = \sqrt{\frac{1}{NK} \sum_{i=1}^{N} \sum_{A=1}^{K} q_{\pi,i}^{(A),\text{naive}}},$$

where $N = 50$ is the number of recordings and $K = 10$ is the number of scorers. These naive-predictor errors are reported as a benchmark for the scale of scorer-versus-panel metric disagreement.

7.2.7 Variance decomposition

To quantify the relative contributions of between-recording differences, scorer-specific effects, and residual variability, we decomposed the variance of each derived sleep metric using a two-way random-effects variance-components model [244], [245]. For metric π, recording i, and scorer A, let

$$Y_{\pi,i}^{(A)} = \Psi_\pi\left(H_i^{(A)}\right)$$

denote the scorer-derived metric value. We modeled this quantity as

$$Y_{\pi,i}^{(A)} = \mu_\pi + u_{\pi,i} + v_{\pi,A} + \varepsilon_{\pi,i,A},$$

where μ_π is the overall mean for metric π, $u_{\pi,i}$ is a random effect for recording i, $v_{\pi,A}$ is a random effect for scorer A, and $\varepsilon_{\pi,i,A}$ is a residual term. The corresponding variance components are

$$u_{\pi,i} \sim (0, \sigma^2_{\pi,\text{rec}}), \quad v_{\pi,A} \sim (0, \sigma^2_{\pi,\text{scorer}}), \quad \text{and} \quad \varepsilon_{\pi,i,A} \sim (0, \sigma^2_{\pi,\text{res}}).$$

For each metric, total variance was defined as

$$\sigma^2_{\pi,\text{tot}} = \sigma^2_{\pi,\text{rec}} + \sigma^2_{\pi,\text{scorer}} + \sigma^2_{\pi,\text{res}}.$$

The percentage of total variance attributable to the recording, scorer, and residual components are thus:

$$100 \times \frac{\sigma^2_{\pi,\text{rec}}}{\sigma^2_{\pi,\text{tot}}}, \qquad 100 \times \frac{\sigma^2_{\pi,\text{scorer}}}{\sigma^2_{\pi,\text{tot}}}, \qquad 100 \times \frac{\sigma^2_{\pi,\text{res}}}{\sigma^2_{\pi,\text{tot}}}.$$

7.2.8 Consensus construction comparison

When multiple independent scorers are available for the same recording, there is more than one way to define a consensus reference for a sleep metric. We therefore considered two conceptually distinct consensus constructions.

For recording i and sleep metric π, metric averaging computes the sleep metric separately from each scorer's hypnogram and then

averages the resulting values. For K scorers, this gives

$$\bar{Y}_{\pi,i}^{(\mathrm{avg})} = \frac{1}{K} \sum_{A=1}^{K} \Psi_\pi\Big(H_i^{(A)}\Big).$$

Second, the majority-vote hypnogram consensus first constructs an epoch-by-epoch consensus hypnogram and then derives the sleep metric from that consensus sequence. Let

$$\hat{p}_{i,t}(s) = \frac{1}{K} \sum_{A=1}^{K} \mathbf{1}\Big\{H_{i,t}^{(A)} = s\Big\}, \qquad s \in S,$$

denote the empirical proportion of scorers assigning sleep stage s at epoch t in recording i. The majority-vote hypnogram is then defined as

$$H_{i,t}^{(\mathrm{MV})} = \arg\max_{s \in S} \hat{p}_{i,t}(s).$$

When two or more sleep stages were tied, ties were resolved deterministically using the predefined sleep-stage-priority order: Wake, N1, N2, N3, and REM.

The corresponding sleep metric is

$$Y_{\pi,i}^{(\mathrm{MV})} = \Psi_\pi\Big(H_i^{(\mathrm{MV})}\Big).$$

These two consensus procedures are not mathematically equivalent, that is, in general

$$\bar{Y}_{\pi,i}^{(\mathrm{avg})} \neq Y_{\pi,i}^{(\mathrm{MV})}.$$

We quantified their difference by computing the absolute difference for each recording i and sleep metric π,

$$\Delta_{\pi,i}^{(\mathrm{cons})} = \Big|\bar{Y}_{\pi,i}^{(\mathrm{avg})} - Y_{\pi,i}^{(\mathrm{MV})}\Big|.$$

These consensus-construction differences were then summarized across recordings for each derived sleep metric.

7.2.9 Bootstrap confidence intervals

Uncertainty in agreement and reliability estimates was quantified using a recording-level nonparametric bootstrap. Each bootstrap sample was constructed by resampling the $N = 50$ recordings

with replacement. For each bootstrap sample, the statistic of interest was recomputed using the same analysis procedure as in the original sample. Confidence intervals were then obtained as 95% percentile bootstrap intervals based on 3000 bootstrap resamples.

7.2.10 Software and implementation

All analyses were implemented in Python 3.11 using custom analysis scripts together with NumPy 2.1.0, pandas 2.2.2, and Matplotlib 3.9.2.

7.3 Results

Figures 7.1 and 7.2 provide a visual summary of the sleep metric variation in the dataset. In Figure 7.1, the 10 scorer-specific hypnograms for one representative recording look broadly similar at first glance: they show the same overall overnight structure, with comparable timing of major sleep cycles. However, closer inspection of the hypnograms shows many small difference. These small differences are not distributed uniformly across sleep stages, and they accumulate into visible spread in the derived sleep metrics shown in the panel on the right, particularly for N1, N2, and N3 duration (minutes), even when the overall hypnogram shape appears similar.

Figure 7.2 shows that this is not just a feature of one illustrative recording. Across the 50 recordings, the within-record spread across scorers is generally narrowest for TST, while NREM sleep-stage-duration metrics show broader and more variable dispersion across recordings. In particular, several recordings show wide scorer-to-scorer spread for N1, N2, and N3, whereas TST is often much more tightly clustered and the REM spread is intermediate. Taken together, these figures suggest that inter-scorer variability arises less from disagreement about the broad overnight structure of sleep and more from how sleep is partitioned into specific sleep stages within that structure.

7.3.1 Epoch-by-epoch staging agreement

Table 7.2 shows epoch-by-epoch staging agreement for the whole dataset. For 5-class sleep staging, Krippendorff's α was 0.784 (95%

Table 7.2: Inter-scorer agreement coefficients for epoch-by-epoch sleep staging.

Measure	Krippendorff's α	95% CI
5-class staging	0.784	[0.763, 0.797]
4-class staging	0.822	[0.808, 0.842]
Wake vs rest	0.911	[0.899, 0.922]
N1 vs rest	0.414	[0.385, 0.428]
N2 vs rest	0.719	[0.699, 0.743]
N3 vs rest	0.755	[0.723, 0.783]
REM vs rest	0.865	[0.857, 0.874]

Krippendorff's α was computed for epoch-by-epoch sleep staging using the nominal difference function.
One-vs-rest rows report agreement for binary nominal staging problems.
Confidence intervals are 95% percentile subject-level bootstrap intervals.
Abbreviations: CI, confidence interval; N1, stage N1 sleep; N2, stage N2 sleep; N3, stage N3 sleep; REM, rapid eye movement sleep.

CI [0.763, 0.797]) . When N1 and N2 were pooled into a 4-class scheme, agreement increased, with Krippendorff's $\alpha = 0.822$ (95% CI [0.808, 0.842]).

In one-vs-rest analyses, Krippendorff's α was highest for Wake vs rest (0.911, 95% CI [0.899, 0.922]) and REM vs rest (0.865, 95% CI [0.857, 0.874]), intermediate for N3 vs rest (0.755, 95% CI [0.723, 0.783]) and N2 vs rest (0.719, 95% CI [0.699, 0.743]), and lowest for N1 vs rest (0.414, 95% CI [0.385, 0.428]).

7.3.2 Agreement of derived sleep metrics

Agreement for derived sleep metrics was strongly metric-dependent (Table 7.3). Metrics that primarily reflected the distinction between sleep and wake showed the highest reliability. By Krippendorff's α, agreement was highest for SOL ($\alpha = 0.982$, 95% CI [0.916, 0.993]), WASO ($\alpha = 0.927$, 95% CI [0.879, 0.952]), SE ($\alpha = 0.903$, 95% CI [0.823, 0.934]), and TST ($\alpha = 0.845$, 95% CI [0.676, 0.898]). REM-related metrics were also relatively reliable, including REM duration (minutes; $\alpha = 0.777$, 95% CI [0.697, 0.825]), REM% ($\alpha = 0.719$, 95% CI [0.654, 0.785]), and ROL (minutes; $\alpha = 0.739$, 95% CI [0.552, 0.838]).

In contrast, metrics describing how total sleep was partitioned across NREM sleep stages showed lower reliability. The weakest

Table 7.3: Inter-scorer agreement coefficients for derived sleep metrics.

Measure	ICC(2,1)	95% CI	ICC(2,10)	95% CI	ICC(3,1)	95% CI	Krippendorff's α	95% CI
Total sleep time (TST; minutes)	0.848	[0.682, 0.900]	0.982	[0.955, 0.989]	0.870	[0.714, 0.922]	0.845	[0.676, 0.898]
N1 duration (minutes)	0.442	[0.384, 0.473]	0.888	[0.862, 0.900]	0.585	[0.534, 0.636]	0.424	[0.361, 0.456]
N2 duration (minutes)	0.615	[0.486, 0.661]	0.941	[0.904, 0.951]	0.744	[0.632, 0.778]	0.604	[0.470, 0.652]
N1/N2 duration (minutes)	0.702	[0.659, 0.749]	0.959	[0.951, 0.968]	0.756	[0.711, 0.794]	0.696	[0.652, 0.745]
N3 duration (minutes)	0.537	[0.391, 0.600]	0.921	[0.864, 0.937]	0.693	[0.571, 0.763]	0.522	[0.366, 0.587]
REM duration (minutes)	0.781	[0.702, 0.828]	0.973	[0.959, 0.980]	0.818	[0.729, 0.854]	0.777	[0.697, 0.825]
N1%	0.517	[0.331, 0.582]	0.915	[0.827, 0.933]	0.646	[0.491, 0.700]	0.503	[0.304, 0.571]
N2%	0.419	[0.273, 0.435]	0.878	[0.789, 0.885]	0.595	[0.444, 0.619]	0.397	[0.239, 0.416]
N1/N2%	0.608	[0.391, 0.667]	0.940	[0.862, 0.952]	0.701	[0.491, 0.752]	0.599	[0.374, 0.659]
N3%	0.540	[0.441, 0.603]	0.922	[0.887, 0.938]	0.699	[0.624, 0.745]	0.525	[0.419, 0.591]
REM%	0.725	[0.662, 0.790]	0.964	[0.951, 0.974]	0.792	[0.734, 0.865]	0.719	[0.654, 0.785]
Sleep onset latency (SOL; minutes)	0.982	[0.917, 0.993]	0.998	[0.991, 0.999]	0.983	[0.924, 0.993]	0.982	[0.916, 0.993]
REM onset latency (ROL; minutes)	0.743	[0.558, 0.840]	0.967	[0.926, 0.981]	0.753	[0.582, 0.852]	0.739	[0.552, 0.838]
Wake after sleep onset (WASO; minutes)	0.928	[0.881, 0.952]	0.992	[0.987, 0.995]	0.937	[0.894, 0.964]	0.927	[0.879, 0.952]
Sleep efficiency (SE; %)	0.904	[0.826, 0.935]	0.990	[0.979, 0.993]	0.919	[0.844, 0.949]	0.903	[0.823, 0.934]

ICC(2,1) corresponds to a two-way random-effects, absolute-agreement, single-rater formulation (ICC(A,1)).
ICC(2,10) corresponds to a two-way random-effects, absolute-agreement, average-rater formulation (ICC(A,10)).
ICC(3,1) corresponds to a two-way mixed-effects, consistency, single-rater formulation (ICC(C,1)).
Krippendorff's α for sleep metrics was computed with the interval difference function.
Confidence intervals are 95% percentile subject-level bootstrap intervals.
Abbreviations: ICC, intraclass correlation coefficient; CI, confidence interval; TST, total sleep time; N1, stage N1 sleep; N2, stage N2 sleep; N3, stage N3 sleep; REM, rapid eye movement; SOL, sleep onset latency; ROL, rapid eye movement onset latency; WASO, wake after sleep onset; SE, sleep efficiency.

agreement was observed for N2% (α = 0.397, 95% CI [0.239, 0.416]) and N1 duration (minutes; α = 0.424, 95% CI [0.361, 0.456]). Reliability was also lower for N1% (α = 0.503, 95% CI [0.304, 0.571]), N3 duration (minutes; α = 0.522, 95% CI [0.366, 0.587]), and N3% (α = 0.525, 95% CI [0.419, 0.591]). N2 duration (minutes; α = 0.604, 95% CI [0.470, 0.652]) and the pooled N1/N2 measures (N1/N2 duration (minutes): α = 0.696, 95% CI [0.652, 0.745]; N1/N2%: α = 0.599, 95% CI [0.374, 0.659]) were intermediate.

The ICC analyses showed the same overall pattern, but from three complementary reliability perspectives (Table 7.3). ICC(2,1), which reflects absolute agreement for a single scorer, was highest for SOL (0.982), WASO (0.928), SE (0.904), and TST (0.848), and was also relatively high for REM-related metrics such as REM duration (minutes; 0.781), REM% (0.725), and ROL (minutes; 0.743). By contrast, ICC(2,1) was substantially lower for several NREM sleep stage metrics, including N1 duration (minutes; 0.442), N2% (0.419), N1% (0.517), N3 duration (minutes; 0.537), and N3% (0.540). ICC(2,10), which reflects the reliability of the mean across all 10 scorers, was uniformly much higher, ranging from 0.878 to 0.998. This indicates that scorer averaging yields a much more reliable reference than any single scorer alone. ICC(3,1), which asks whether different scorers rank subjects similarly, even if some scorers are systematically higher or lower than others, was generally somewhat higher than ICC(2,1), especially for metrics such as N1 duration (minutes), N2%, and N3 duration (minutes), but the overall ranking of metrics remained the same. Thus, the ICC results reinforce the conclusion that reliability depends strongly on which sleep metric is being considered, and that multi-scorer averaging materially improves the stability of the reference value.

7.3.3 Leave-one-scorer-out error

Leave-one-scorer-out error varied by metric (Table 7.4). The last column reports the corresponding average error of a naive constant predictor for each metric, where available. For robust metrics such as TST, REM duration (minutes), REM%, WASO, and SE, scorer-versus-panel errors were far smaller than the naive baseline. In contrast, for several weaker sleep-stage-composition metrics, the separation from the naive baseline was much smaller. Notably, for N1 duration (minutes), N1%, N2%, N1/N2%, and N3%, the

Table 7.4: Leave-one-scorer-out error and comparison with a naive constant predictor.

Metric	Avg. MAE	MAE range	Avg. RMSE	RMSE range	Naive Avg. MAE / RMSE
TST	14.2	[8.8, 22.6]	25.1	[16.2, 33.6]	56.6 / 69.4
N1 duration (minutes)	13.8	[8.1, 25.7]	18.2	[11.2, 32.7]	**16.7** / **23.3**
N2 duration (minutes)	27.1	[16.6, 44.0]	34.0	[22.1, 52.1]	**43.0** / 53.1
N1/N2 duration (minutes)	23.9	[16.4, 33.5]	32.1	[20.8, 47.4]	45.5 / 57.1
N3 duration (minutes)	22.7	[13.6, 46.7]	28.6	[17.3, 51.3]	**34.5** / **43.3**
REM duration (minutes)	9.3	[7.4, 17.3]	12.7	[9.3, 22.7]	23.9 / 30.8
N1%	3.2	[1.9, 5.2]	4.3	[2.8, 6.8]	**4.0** / **5.9**
N2%	6.3	[3.4, 9.8]	7.7	[4.2, 11.6]	**8.3** / **10.2**
N1/N2%	5.2	[3.5, 8.5]	6.6	[4.2, 10.3]	**8.4** / 10.4
N3%	5.3	[3.2, 10.6]	6.6	[4.0, 11.8]	**7.8** / **9.6**
REM%	2.3	[1.6, 3.8]	2.9	[2.1, 4.7]	4.6 / 5.8
SOL (minutes)	5.7	[4.1, 7.0]	12.7	[7.7, 21.6]	— / —
ROL (minutes)	14.4	[10.0, 16.8]	29.6	[16.5, 37.0]	— / —
WASO (minutes)	12.4	[7.5, 20.8]	21.6	[12.8, 31.2]	77.2 / 103.4
SE (%)	2.3	[1.4, 3.6]	4.0	[2.6, 5.4]	18.1 / 21.5

Avg. MAE and Avg. RMSE denote mean scorer-versus-leave-one-out-consensus error across scorers.
Ranges denote the minimum and maximum scorer-specific mean errors across the 10 scorers.
The last column reports the corresponding average error of a naive constant predictor for each metric, where available.
Bold values indicate that the naive average error falls within the corresponding human scorer range for that metric.
Abbreviations: Avg., average; MAE, mean absolute error; RMSE, root mean squared error; TST, total sleep time; N1, stage N1 sleep; N2, stage N2 sleep; N3, stage N3 sleep; REM, rapid eye movement; SOL, sleep onset latency; ROL, rapid eye movement onset latency; WASO, wake after sleep onset; SE, sleep efficiency.

naive average error fell within the observed human scorer range, indicating that scorer-versus-panel disagreement for these metrics approached the scale of a simple population-mean predictor on average.

Figure 7.3 shows a generally increasing relationship between leave-one-out staging disagreement and downstream metric error. However, the relationship differed between metrics. For TST, SE, and WASO, scorer-versus-panel errors were usually well below the naive benchmark. For N1-, N2-, and N3-related composition metrics, especially N1 duration (minutes), N1%, N2 duration (minutes), N2%, N3 duration (minutes), and N3%, this was less consistently the case, and many points approached or exceeded the naive line at higher disagreement levels. SOL and especially ROL were mostly near zero but showed occasional large outliers.

7.3.4 Variance decomposition

Variance decomposition showed that the more robust metrics (TST, SE, SOL, ROL, and WASO) were dominated by between-recording variance, whereas weaker metrics (NREM duration (minutes) and percentages) showed larger scorer and residual components, with REM duration (minutes) and percentage showing intermediate between-recording variance (Table 7.5). These results indicate that disagreement is concentrated less in sleep–wake continuity measures, and more in how sleep is partitioned across NREM sleep stages.

7.3.5 Consensus construction comparison

We examined whether different approaches to consensus construction yielded the same sleep metric values. For each recording, we compared two possible reference definitions: (1) averaging the metric values computed separately from each scorer's hypnogram and (2) deriving the metric from an epoch-by-epoch majority-vote hypnogram. In Figures 7.1 and 7.2, these two consensus constructions are shown directly on the metric plots: the metric-averaged consensus is shown by the black horizontal line, and the majority-vote-derived consensus is shown by the white triangle.

These two procedures did not, in general, produce the same result. In the example recording shown in Figure 7.1, metric averaging differed from majority-vote-derived values by -0.50

minutes for TST, −13.65 minutes for N1/N2, +0.85 minutes for N3, and +12.30 minutes for REM, where negative values indicate lower values under metric averaging.

Across all 50 recordings (Table 7.6), the absolute difference between the metric-averaged consensus and the majority-vote-derived consensus was 5.6 ± 7.3 minutes for TST, 6.2 ± 4.1 minutes for N1 duration, 10.2 ± 9.5 minutes for N2 duration, 7.5 ± 7.1 minutes for N1/N2 duration, 7.0 ± 5.8 minutes for N3 duration, and 3.0 ± 3.3 minutes for REM duration. For wake-vs-sleep metrics, the corresponding differences were 3.4 ± 6.0 minutes for SOL, 10.4 ± 24.1 minutes for ROL, and 4.6 ± 6.2 minutes for WASO (mean ± SD).

These results show that the definition of consensus can have an effect on the sleep metrics, most noticeably N2 duration (minutes) and ROL.

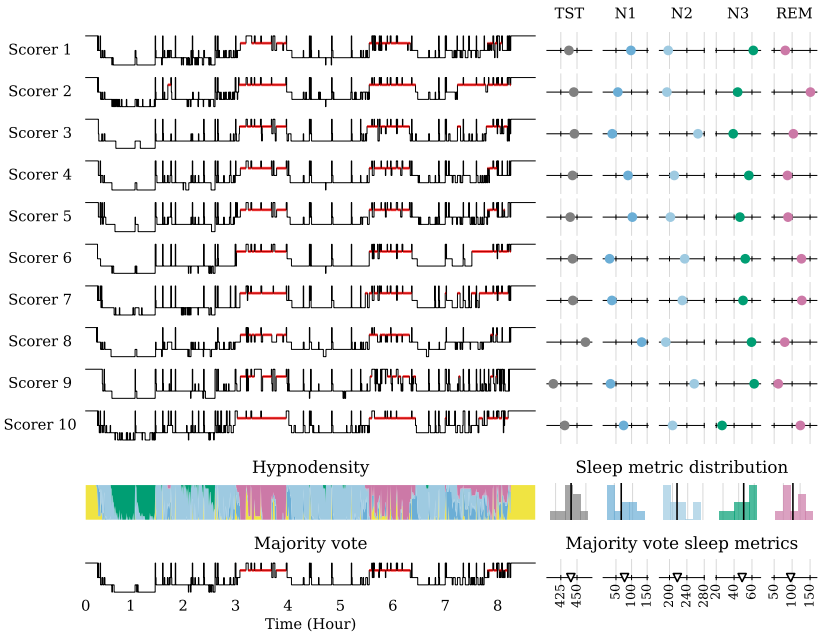


Figure 7.1: Hypnograms and derived sleep architecture metrics for a single subject-night scored independently by 10 expert scorers. Rows show individual scorers' hypnograms (left) and the corresponding derived sleep metrics (right). The empirical epoch-by-epoch sleep stage distribution across scorers is shown as an aggregated probability plot below the scorer hypnograms, and the majority-vote hypnogram is shown at the bottom. Distributions of scorer-derived sleep metrics are shown alongside the metric values derived from the majority-vote hypnogram (triangle) and the mean of the scorer-derived values (black vertical line), in minutes. Abbreviations: TST, total sleep time; N1, stage N1 sleep; N2, stage N2 sleep; N3, stage N3 sleep; REM, rapid eye movement sleep.

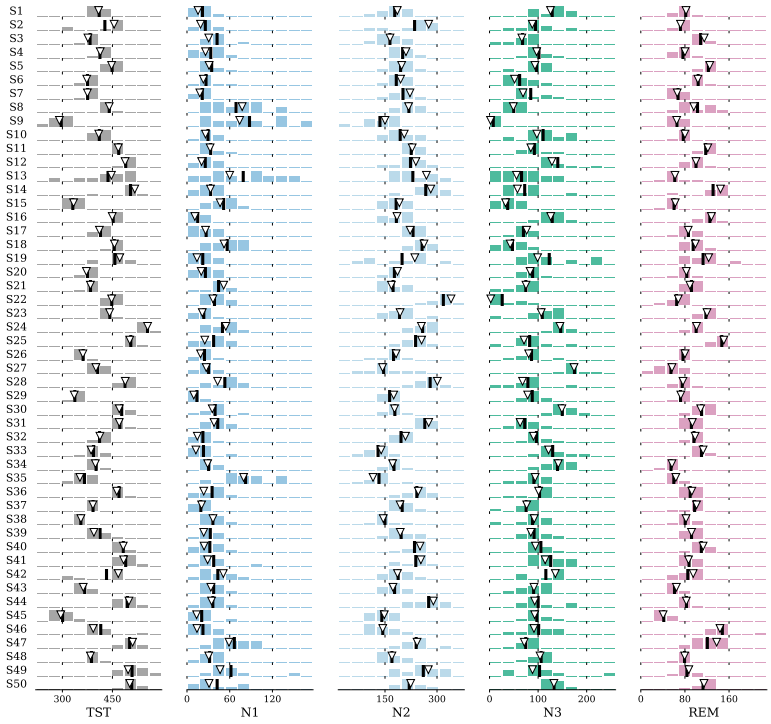


Figure 7.2: Distribution of scorer-derived sleep architecture metrics across all 50 recordings in the Sleep Revolution Multicenter Scoring dataset. Each cell shows a histogram of scorer-derived values for one metric in one recording. Columns correspond to TST, N1/N2 duration (minutes), N3 duration (minutes), and REM duration (minutes). The black vertical line marks the value derived from the majority-vote hypnogram, and the white triangle with black border marks the mean of the scorer-derived metric values. Abbreviations: TST, total sleep time; N1/N2, pooled stage N1 and stage N2 sleep; N3, stage N3 sleep; REM, rapid eye movement sleep.

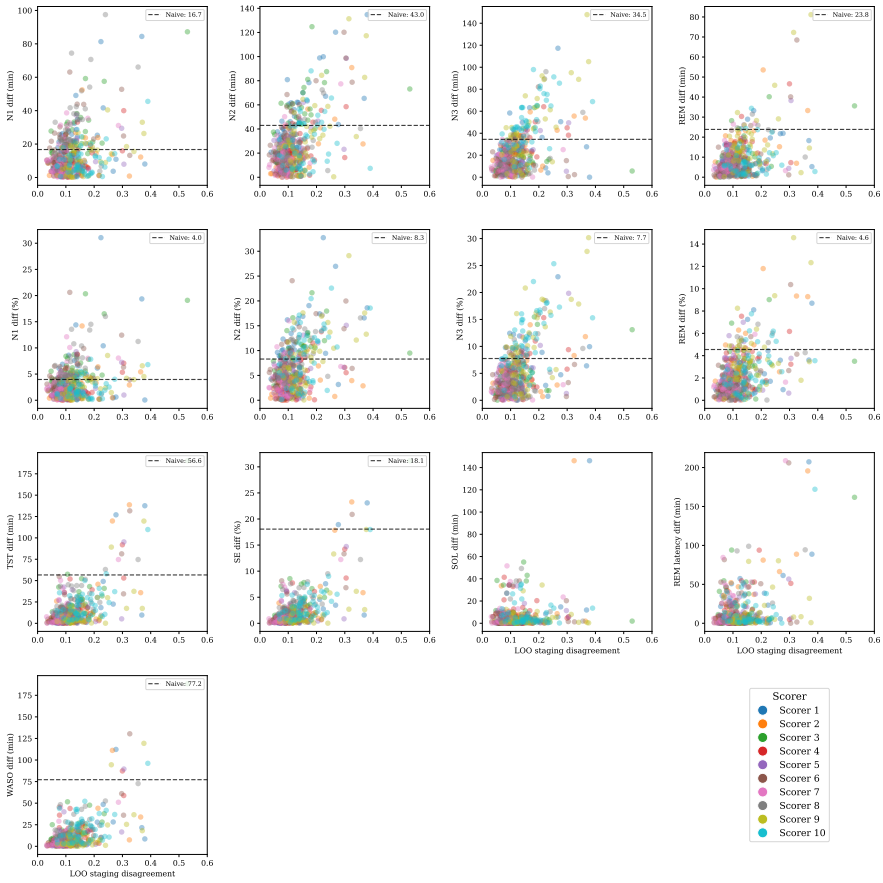


Figure 7.3: Relationship between leave-one-out epoch-by-epoch staging disagreement and downstream error in derived sleep architecture metrics. Each point represents one scorer–recording pair. The x-axis shows leave-one-out staging disagreement, and the y-axis shows the absolute difference between the scorer-derived metric and the corresponding leave-one-out (LOO) panel mean for the displayed metric. Colors identify scorers. Dashed horizontal lines denote the average absolute error of a naive constant predictor for metrics where such a benchmark was available. Abbreviations: LOO, leave-one-out; diff., difference; TST, total sleep time; SE, sleep efficiency; SOL, sleep onset latency; WASO, wake after sleep onset; N1, stage N1 sleep; N2, stage N2 sleep; N3, stage N3 sleep; REM, rapid eye movement.

Table 7.5: Variance decomposition of derived sleep metrics across recordings and scorers.

	Variance component (%)		
Metric	**Recording**	**Scorer**	**Residual**
TST	84.8	2.5	12.6
N1 duration (minutes)	44.2	24.4	31.4
N2 duration (minutes)	61.5	17.3	21.2
N1/N2 duration (minutes)	70.2	7.2	22.7
N3 duration (minutes)	53.7	22.5	23.8
REM duration (minutes)	78.1	4.5	17.4
N1%	51.7	20.0	28.3
N2%	41.9	29.6	28.5
N1/N2%	60.9	13.2	26.0
N3%	54.0	22.7	23.3
REM%	72.5	8.5	19.0
SOL (minutes)	98.2	0.1	1.7
ROL (minutes)	74.3	1.3	24.4
WASO (minutes)	92.8	0.9	6.2
SE (%)	90.4	1.6	8.0

Variance was decomposed into recording, scorer, and residual components using the two-way random-effects variance-components model described in the Methods.

Percentages are proportions of total variance.

Abbreviations: TST, total sleep time; N1, stage N1 sleep; N2, stage N2 sleep; N3, stage N3 sleep; REM, rapid eye movement; SOL, sleep onset latency; ROL, rapid eye movement onset latency; WASO, wake after sleep onset; SE, sleep efficiency.

Table 7.6: Differences between two consensus constructions for hypnogram-derived sleep architecture metrics in the Sleep Revolution Multicenter Scoring dataset.

Metric	Mean absolute difference ± SD (min)
Total sleep time	5.6 ± 7.3
N1 duration	6.2 ± 4.1
N2 duration	10.2 ± 9.5
N1/N2 duration	7.6 ± 7.1
N3 duration	7.0 ± 5.8
REM duration	3.0 ± 3.3
Sleep onset latency	3.4 ± 6.0
REM onset latency	10.4 ± 24.1
Wake after sleep onset	4.6 ± 6.2

The table compares two consensus constructions: (1) averaging metric values computed separately from each scorer's hypnogram and (2) deriving metric values from an epoch-by-epoch majority-vote hypnogram. Values are reported as mean absolute difference ± standard deviation, in minutes.

Abbreviations: SD, standard deviation; N1, stage N1 sleep; N2, stage N2 sleep; N3, stage N3 sleep; REM, rapid eye movement.

7.4 Discussion

Epoch-by-epoch staging agreement was high overall and improved further when N1 and N2 were pooled, with the strongest one-vs-rest agreement for Wake and REM and the weakest for N1. At the level of derived sleep metrics, reliability varied markedly by metric: SOL, WASO, SE, TST, and REM-related measures showed the strongest agreement, whereas metrics describing the partitioning of sleep across NREM sleep stages, especially N1-, N2-, and N3-related measures, were clearly less reliable. These results show that strong agreement at the epoch-by-epoch level does not translate uniformly into strong agreement for downstream sleep architecture metrics. A clear theme running through the results was that hypnogram-derived sleep metrics did not behave as a single class, but instead fell into three broad groups. The first comprised wake–sleep metrics (SOL, WASO, SE, and TST), the second comprised REM-related metrics (REM duration, REM%, and ROL), and the third comprised NREM sleep-stage-composition metrics (N1, N2, N1/N2, and N3 duration and percentage measures). The distinction between these three groups of sleep metrics recurred across the reliability coefficients, scorer-versus-panel errors, and variance components. This pattern was reinforced by the leave-one-out analyses and variance decomposition. Scorer-versus-panel errors were smallest for SE, REM%, and SOL, but were much larger for duration-based NREM metrics such as N2, N1/N2, and N3 duration (minutes). For several sleep-stage-composition metrics, scorer disagreement approached the scale of a naive constant predictor, and variance decomposition showed larger scorer-dependent and residual components for these measures than for more robust metrics. Finally, different consensus constructions did not generally yield the same metric values, with especially notable differences for N2 duration (minutes) and ROL.

Our first research question asked how variable commonly used sleep architecture metrics are across expert scorers. The answer was clearly metric-specific. Using the conventional ICC interpretation of < 0.50 poor, 0.50–0.75 moderate, 0.75–0.90 good, and > 0.90 excellent [246], single-scorer absolute-agreement reliability (ICC(2,1)) was *excellent* for SOL (0.982), WASO (0.928), and SE (0.904); *good* for TST (0.848) and REM duration (minutes; 0.781); *moderate* for ROL (0.743), REM% (0.725), N1/N2 duration (minutes; 0.702), N2 duration (minutes; 0.604), N1/N2% (0.608), N3

duration (minutes; 0.537), N3% (0.540), and N1% (0.517); and *poor* for N1 duration (minutes; 0.442) and N2% (0.419).

By Krippendorff's α, where values below 0.67 are commonly treated as not adequate to support a conclusion and values between 0.67 and 0.80 as supporting only tentative conclusions [242], [248], only SOL (0.982), WASO (0.927), SE (0.903), and TST (0.845) were clearly above 0.80. REM duration (minutes; 0.777), REM% (0.719), ROL (0.739), and N1/N2 duration (minutes; 0.696) fell into the tentative range. All remaining sleep-stage-partition metrics were below 0.67, including N1 duration (minutes; 0.424), N2 duration (minutes; 0.604), N3 duration (minutes; 0.522), N1% (0.503), N2% (0.397), N1/N2% (0.599), and N3% (0.525).

Thus, the most dependable derived metrics were the sleep–wake continuity measures and TST, with REM duration (minutes) also performing relatively well. By contrast, the weakest metrics were those describing how total sleep was partitioned across NREM sleep stages, especially N1 and N2. N3 was not among the very weakest metrics by ICC(2,1), but neither N3 duration (minutes) nor N3% reached the usual Krippendorff threshold for dependable conclusions. Notably, when the 10 scorers were averaged, ICC(2,10) rose to the good-to-excellent range for all metrics, showing that multi-scorer averaging yields a much more reliable reference than any single scorer alone.

Perhaps the clearest sign of poor inter-scorer reliability is that, for several of the weaker metrics, some individual scorers were, on average, no better than a naive constant predictor. This was the case for all NREM duration (minutes) and NREM% metrics, as well as the pooled N1/N2% metric, in which the naive benchmark fell within the observed human scorer range.

Our second research question concerned how disagreement in epoch-by-epoch staging relates to disagreement in derived sleep metrics. At the epoch level, agreement was fairly good overall: 5-class and pooled 4-class staging both showed reasonably strong agreement, Wake and REM were especially robust, N1 was the weakest sleep stage, and N3 agreement was intermediate. However, agreement at the label level and agreement at the metric level are not the same thing.

The key conceptual point is that sleep metrics are aggregated quantities. Small disagreements in sleep stage allocation can accumulate across a night into materially different summary values, especially when those disagreements affect transitions between

adjacent NREM sleep stages or the boundaries of sleep stage segments. This helps explain why hypnograms that appear broadly similar on visual inspection can nonetheless yield noticeably different sleep metrics. In that sense, visually similar staging solutions do not necessarily imply equivalent downstream interpretations.

The practical implication is that downstream clinical or research interpretation should focus on the stability of the metric itself, not only on staging agreement. Epoch-by-epoch agreement remains important, but it is not a sufficient proxy for the dependability of every hypnogram-derived metric.

For most of the stronger metrics, variance was driven more by between-recording differences than by who did the scoring. Variance decomposition then suggested an important difference between the REM- and NREM-related metrics. For REM duration (minutes), REM% and ROL, scorer-specific variance was relatively small, and the remaining non-recording variability was driven more by residual variance. By contrast, several NREM sleep-stage-composition metrics showed substantial scorer-specific variance in addition to sizable residual variance, indicating greater dependence on who performed the scoring. One way to explain this difference is that disagreement in the REM-related metrics may be driven more by data ambiguity, whereas disagreement in the NREM metrics may reflect differences among scorers in how they applied the scoring rules.

N3 sleep is often treated as a particularly important component of sleep architecture and differences in N3 composition are frequently interpreted as biologically or clinically informative. In the data, however, both N3 duration (minutes) and N3% showed only moderate reliability, substantial scorer-dependent variance, and sizeable scorer-versus-panel errors. For both N3 duration (minutes) and N3%, the naive constant predictor's average error fell within the observed range of mean human scorer errors. Numerous studies have linked N3 sleep to health-related outcomes, including metabolic health, cardiovascular outcomes, and brain aging, often using N3 metrics derived from single-scorer PSG staging [249]–[252]. Our findings suggest that future studies using N3 as an outcome or explanatory variable should not rely on single-scorer PSG staging.

When multiple independent scorers are available, the choice of reference construction matters. In our data, averaging the scorer-derived metric values and deriving the metric from a majority-vote

hypnogram did not, in general, give the same result. Because these approaches give different values, the reference construction should be chosen with the quantity being estimated in mind.

We argue that for sleep metrics, averaging scorer-specific metric values is the more appropriate reference. It represents expert opinion directly at the level of the quantity being analyzed, rather than after reducing the data to a single hypnogram. It also avoids artifacts introduced by epoch-by-epoch voting and tie-breaking, which can affect downstream metric values without reflecting the central tendency of expert opinion at the metric level.

These results have direct implications for how novel alternative sleep measurement methods, such as automated sleep staging systems and wearable sleep-monitoring technologies, should be evaluated. If the intended goal is to estimate sleep architecture, then evaluation should be performed at the level of the derived sleep metrics themselves, not only at the level of epoch-by-epoch agreement. Epoch-by-epoch agreement remains important, but the present results show that it does not map well onto agreement in downstream sleep metrics.

Our results also show that any such evaluation should rely on multiple expert scorers rather than a single human reference. Several of the sleep metrics examined varied meaningfully across scorers, suggesting that the reference value for a given recording is not well represented by a single scorer. Evaluation against a single scorer, therefore, risks conflating model error with ordinary expert variability.

When a consensus reference is required, the results further suggest that it should be constructed by averaging the scorer-specific metric values, rather than by first forming a majority-vote hypnogram and then deriving the metric from that hypnogram. These two procedures did not, in general, yield the same result, and metric averaging is the more direct way to represent panel opinion at the level actually being evaluated.

This study has several strengths. It analyzed 10 independent scorers evaluating the same 50 full-night PSG recordings under a shared scoring framework, allowing direct quantification of inter-scorer variability at the metric level. It also used several complementary analytical perspectives, including ICCs, Krippendorff's α, leave-one-scorer-out error, naive benchmark comparison, and variance decomposition. Together, these provided a more complete picture of reliability than any single metric alone.

The study also has important limitations. First, it included only 50 PSG recordings. Second, it was based on a single dataset and a single common scoring framework, so the observed reliability structure may not generalize unchanged to other datasets, scorer panels, sleep laboratories, or scoring conventions. Because of this, our results likely overestimate the reliability of sleep metrics across scorers in typical clinical conditions. Third, the analysis quantifies inter-scorer variability, not physiological truth, and not direct clinical validity. A stable metric is not necessarily the most clinically meaningful one, and an unstable metric is not necessarily clinically irrelevant. Fourth, no explicit lights-off or lights-on markers were available, so TIB and, therefore, SE and SOL were based on the full recording duration rather than on true in-bed intervals. These limitations do not undermine our main results, but they do define the scope.

Future work is needed. First, the same analyses should be replicated in other PSG datasets and scorer panels. Second, it would be valuable to test whether the same reliability pattern holds across diagnostic subgroups, since scorer uncertainty may differ by sleep disorders and subject characteristics. Third, some sleep metrics may be better represented as intervals or scorer distributions rather than single-point targets, particularly when inter-scorer variability is substantial. Fourth, evaluation frameworks for automated sleep analysis and wearable sleep trackers should explicitly account for scorer uncertainty. Fifth, the reliability of N3 sleep metrics warrant focused investigation, because they are clinically important yet only moderately dependable in the present data.

Finally, we suggest that automated methods for sleep architecture analysis should treat sleep architecture metrics as estimation targets rather than by-products of a staged hypnogram.

7.5 Conclusion

In this study, we showed that inter-scorer reliability in sleep architecture is strongly metric-dependent. Reasonably strong agreement at the epoch-by-epoch level did not translate uniformly into reliable downstream sleep metrics. Sleep-wake continuity measures such as SOL, WASO, SE, and TST were the most reliable, REM-related metrics were intermediate, and metrics describing how sleep is partitioned across NREM sleep stages were substantially

less reliable. Hypnogram-derived sleep metrics should therefore not be treated as equally dependable reference quantities.

These findings have practical implications for both sleep research and sleep medicine. If the goal is to estimate sleep architecture, evaluation should be performed at the level of the derived sleep metrics themselves, not only at the level of epoch-by-epoch staging agreement. Such evaluation should also rely on multiple expert scorers rather than a single human reference, because metric values can vary substantially across scorers. Our results indicate that when a consensus reference is needed, averaging scorer-specific metric values is more appropriate than deriving metrics from a majority-vote hypnogram.

Chapter 8

Rethinking prediction targets in automated sleep analysis

From sleep-stage classification to sleep-metric estimation

Full reference: (Manuscript in preparation.) E. Hardarson *et al.*, 2026, "From sleep staging to sleep metrics: Rethinking targets in automated sleep analysis"

8.1 Introduction

Sleep architecture metrics such as total sleep time (TST) and time spent in different sleep stages are derived from overnight polysomnography (PSG) and are widely used in clinical sleep medicine and sleep research. Under current clinical guidelines [8], sleep metrics are derived from a multi-step workflow as shown in Figure 8.1. First, PSGs are divided into 30-second segments (called *epochs*), and each epoch is manually classified by an expert scorer into one of five sleep stages (Wake, rapid eye movement (REM) sleep or non-rapid eye movement sleep stages N1, N2 and N3), following a set of rules defined by the American Academy

of Sleep Medicine (AASM) [8]. The resulting sequence of sleep stage labels, known as the *hypnogram*, is then used to calculate summary sleep metrics.

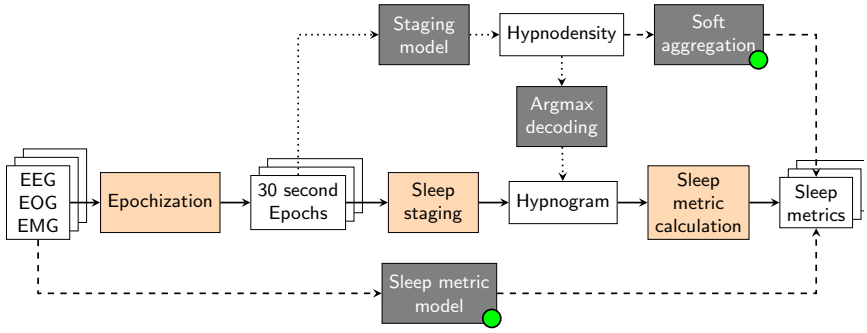


Figure 8.1: A conceptual overview of the clinical sleep diagnosis workflow. Biosignal data from polysomnography recordings are segmented into 30-second epochs. The epochs are annotated by human experts or automatic staging models to produce a sequence of sleep stages called a hypnogram. Sleep metrics, such as the amount of time spent in each sleep stage, are then derived from the hypnogram. This chapter considers two ways to bypass the hypnogram. First, when a staging model outputs epoch-by-epoch sleep stage probabilities (a hypnodensity), sleep metrics can be computed probabilistically by operating on the hypnodensity (soft aggregation). Second, sleep metrics can be estimated directly using a dedicated model. The dotted line indicates the conventional sleep staging machine learning pipeline, while the dashed lines indicate the proposed pathways of analysis which bypass the hypnogram. Abbreviations: EEG, electroencephalography; EOG, electrooculography; EMG, electromyography.

This two-step process (scoring of sleep stages, followed by calculation of sleep metrics) has been the standard approach in sleep analysis since the late 1960s [14]. Manual epoch-by-epoch classification of sleep stages is resource-intensive, requiring an estimated 1.5-2.0 hours of work on average per PSG [12]. The process is also subject to inter- and intra-scorer variability: on average, two human scorers will agree on the sleep stage of an epoch 80–87% of the time [12], [194] and a single scorer will repeat the same staging decision 85–90% of the time during repeated scoring [26]. Numerous automated sleep staging methods have been proposed,

many of which use machine learning models to classify sleep stages with accuracy comparable to inter-scorer agreement [26]. However, existing automated PSG-based sleep staging methods follow the same two-step process as manual scoring: first classifying sleep stages and then calculating sleep metrics [253], [254]. Our earlier work showed that agreement in epoch-by-epoch staging does not translate uniformly into agreement in downstream sleep metrics, and that the reliability of hypnogram-derived sleep metrics is strongly metric-dependent (see Chapters 6 and 7). This makes the conventional hypnogram pipeline a potentially unstable intermediate representation for sleep metric estimation. Figure 8.1 illustrates this target mismatch: conventional automated sleep analysis maps PSG signals to sleep stage labels and only then to sleep metrics, even when the final quantities of interest are the metrics themselves.

One attempt to relax this rigid epoch-hypnogram formulation in the literature is the *hypnodensity* approach. First introduced by Stephansen *et al.* [255], hypnodensity represents each epoch not as a single sleep stage but as a probability distribution over possible sleep stages, thereby conveying more information about sleep trends. Subsequent work has shown that hypnodensity can help identify ambiguous epochs with high manual inter-scorer variability [111] and provide a richer description of sleep architecture than hypnograms [195], [196]. However, the hypnodensity approach still treats epoch-by-epoch sleep staging as an intermediate representation from which sleep metrics are derived.

Prior research has also proposed methods for estimating sleep metrics directly from PSG without relying on traditional sleep stage classification, but these have been mostly limited to wearable sensor data such as actigraphy. For example, Van Hees et al. [256], [257] directly estimated metrics such as TST and sleep onset latency from wrist accelerometry. Although these methods are not PSG-based, they illustrate that sleep metrics can be treated as primary prediction targets rather than as downstream outputs of sleep stage classification.

More broadly, related work has proposed alternative physiological descriptions of sleep that do not depend entirely on conventional sleep stage labels [258]–[265].

Despite these developments, most automated PSG systems are still optimized around the hypnogram, even when the outputs of interest are the derived sleep metrics. This raises a practical

question for automated sleep analysis: when the endpoint is a sleep metric, is the hypnogram a necessary intermediate representation, or does it introduce avoidable error?

In this chapter, we investigate whether sleep metrics can be estimated more robustly from PSG by reducing or bypassing reliance on the hypnogram representation. To this end, we compare three approaches with progressively less reliance on the hypnogram: (1) soft aggregation of hypnodensities produced by existing sleep staging models, which preserves probabilistic sleep stage information when deriving sleep metrics; (2) a deep learning model trained to predict sleep metrics directly from full-night PSG spectrograms, treating the metrics themselves as the prediction targets; and (3) a heuristic model that estimates sleep metrics from physiologically motivated spectral signal characteristics rather than explicit sleep stage labels. We train and evaluate these approaches on a large PSG dataset and benchmark them on a separate multicenter dataset scored independently by ten human experts. Rather than treating sleep stage classification as the primary endpoint, we evaluate these approaches in terms of the sleep metrics generated.

8.2 Methods

8.2.1 Datasets

The analysis in this work used two datasets: the Sleep Heart Health Study dataset (SHHS) and the Sleep Revolution Multicenter Scoring dataset (SRMS). Both datasets contain PSGs with epoch-by-epoch manual sleep stage labels. We used SHHS for training models and in-distribution evaluation and SRMS for quantifying inter-scorer sleep metric variability and benchmarking models against human scorers in a leave-one-out consensus reference.

The SHHS data set was collected for a prospective cohort study designed to investigate obstructive sleep apnea (OSA) as a risk factor for cardiovascular disease. The participants had a type II PSG, performed at home after setup by a sleep technologist [157], [158], [266]. Overall, 6441 participants were recruited. The SHHS dataset was divided into a training set and a hold-out test set, with 20% of the subjects retained for final testing. During training, we employed a K-fold cross-validation procedure on the training split.

For evaluating the performance of models, we used the SRMS,

Table 8.1: Cohort characteristics for the datasets used in this study (SHHS and Sleep Revolution). Abbreviations: SHHS, Sleep Heart Health Study; BMI, body mass index.

	SHHS	Sleep Revolution
Total number of subjects	6441	50
Male/female subjects	3035/3406	29/21
Age, years (mean)	63	42.9
BMI, kg/m^2 (mean)	28.5	27.3
Total sleep time, minutes (mean)	422	426.3

a dataset designed to assess multicenter sleep stage scoring agreement [17]. This dataset comprises 50 Type II PSG performed at home after setup by a sleep technologist, recordings collected from February to June 2021 using Nox A1 devices (Nox Medical, Reykjavik, Iceland). Each recording was independently scored by ten experienced scorers from seven independent sleep centers in accordance with the AASM scoring rules (version 2.6). The population of the study was stratified into the following groups: OSA risk group, restless legs syndrome (RLS) risk group, insomnia risk group, and healthy individuals. The following questionnaires were used to assess risk: the STOP-BANG questionnaire [267] for OSA risk, the Insomnia Severity Index [268] for insomnia risk, and the International Restless Legs Syndrome for RLS [269] risk. Table 8.1 summarizes cohort demographics.

8.2.2 Hypnogram-based sleep metrics

Under the conventional workflow based on AASM scoring rules, sleep metrics are computed from PSG signals via the hypnogram. For a single overnight recording, let

$$X_{1:T} = (X_t)_{t=1}^{T}, \qquad X_t \in \mathbb{R}^C,$$

denote the multichannel time series (in total C channels, e.g., electroencephalography (EEG), electrooculography (EOG), and electromyography (EMG), sampled at epoch length Δt, typically 30 seconds. The "lights-off" and "lights-on" times are denoted as t_0 and t_{end}, respectively.

A staging procedure (automatic or human) produces a hypno-

gram $H^{(A)} := \left(H_t^{(A)}\right)_{t=1}^{T} \in \mathcal{S}^T$ from the signals X, and reported sleep parameters

$$\widetilde{Y}_{\pi}^{(A)} \;=\; \Psi_{\pi}\big(H^{(A)}\big) \;\in\; \mathbb{R}, \qquad \pi \in \Pi,$$

where Ψ_π is the mapping from hypnogram to sleep metric π, as defined by the scoring rules, and $\mathcal{S} = \{\mathrm{W}, \mathrm{N1}, \mathrm{N2}, \mathrm{N3}, \mathrm{REM}\}$ is the space of discrete sleep states, and with $\mathcal{S}_{\mathrm{sleep}} = \mathcal{S} \setminus \{\mathrm{W}\}$.

In this work, we focus on additive duration metrics (minutes), which can be expressed by counting epochs assigned to a sleep stage (or set of sleep stages) and multiplying by Δt. Total sleep time (TST) is defined as the total time spent in non-wake epochs, and sleep stage durations are defined analogously. We have

$$\Psi_{\mathrm{TST}}(s_{1:T}) = \Delta t \sum_{t=1}^{T} \mathbf{1}\{s_t \in \mathcal{S}_{\mathrm{sleep}}\},$$

$$\Psi_{\mathrm{Dur}}(s; s_{1:T}) = \Delta t \sum_{t=1}^{T} \mathbf{1}\{s_t = s\},\ s \in \mathcal{S},$$

where $\mathbf{1}\{\cdot\}$ is the indicator function (1 if the condition holds, 0 otherwise). When reporting combined light sleep (N1 and N2), we use the pooled duration $\Psi_{\mathrm{N1/N2}}(s_{1:T}) = \Delta t \sum_{t=1}^{T} \mathbf{1}\{s_t \in \{\mathrm{N1}, \mathrm{N2}\}\}$. The sleep metrics we evaluate are TST, pooled light sleep duration (N1/N2), N3 duration, and REM duration.

Because hypnogram-derived sleep metrics vary across expert scorers, a single scorer-derived value cannot be treated as a definitive ground truth. When multiple scorers are available, the closest observable proxy is therefore a consensus reference. In this work, we define that reference at the sleep-metric level by averaging the metric values derived independently by each scorer, rather than by deriving sleep metrics from a majority-vote hypnogram, since these two constructions are not equivalent. When comparing an individual scorer to the consensus, we use a leave-one-out (LOO) reference in which that scorer is excluded from the consensus calculation. Thus, for each scorer-versus-consensus comparison on SRMS, the reference value is the mean of the corresponding sleep metric across the other nine scorers. The same LOO metric-mean consensus was used as the reference for model-versus-consensus comparisons on SRMS.

8.2.3 Estimation families

Sleep metric estimators can be grouped according to the intermediate representation they use between the observed PSG signals X and the final sleep metric estimate. In this work, we consider three families: (1) estimation via a discrete hypnogram, (2) estimation via hypnodensity, and (3) direct estimation from the PSG signals. This organization distinguishes conventional staging-based pipelines from approaches that preserve epoch-by-epoch uncertainty and from approaches that bypass staging altogether.

(1) Estimation via hypnogram $X_t\,(\to q_t) \to H_t \to Y$ In the conventional workflow, an estimator produces a discrete hypnogram $\hat{H}_{1:T} \in \mathcal{S}^T$ (at 30 s epochs). This may be obtained directly, or via intermediate epoch-by-epoch posteriors $q_t(s) = P(S_t = s \mid X)$. These posteriors, termed *hypnodensity*, are a probabilistic hypnogram. In the conventional workflow, sleep metrics are computed by applying $\Psi_\pi(\,\cdot\,)$ to the estimated hypnogram.

$$\hat{H}_{1:T} \in \arg\max_{s_{1:T} \in \mathcal{H}^T} \mathbb{P}(H_{1:T} = s_{1:T} \mid X), \quad \widehat{Y}^{(1)} = \Psi\big(\hat{H}_{1:T}\big).$$

(2) Estimation via hypnodensity $X_t \to q_t \to Y$ In this family of sleep metric estimation, the estimator maps the hypnodensity, $\hat{q}_t$ directly to parameters:

$$\widehat{Y}^{(2)} = \Psi\big(q_{1:T}\big).$$

For *additive* metrics, Ψ_π depends only on epoch-by-epoch marginals, so $\Psi_\pi(\hat{q}_{1:T})$ is found by computing the conditional expectation of the additive function: If $U(S) = \sum_{t=1}^{T} u_t(S_t)$, then $\mathbb{E}[U(S) \mid X] = \sum_{t=1}^{T} \sum_{s \in \mathcal{S}} u_t(s)\, \hat{q}_t(s)$. For *path-dependent* metrics, such as sleep onset latency and other run-lengths, Ψ_π is not determined by marginals alone. In that case, one can determine Ψ_π by incorporating transition information $\hat{q}_{t,t+1}(s, s')$. Alternatively, both additive and path-dependent metrics can be approximated by sampling hypnogram paths $S_{1:T}^{(b)} \sim P(S \,|\, X)$ and averaging $\Psi(S^{(b)})$.

(3) Direct estimation $X_t \to Y$ The third family of estimation methods estimates sleep metrics end-to-end. An estimator, f_θ

maps the full-night signals (or a representation of them such as a spectrogram) directly to a vector of sleep estimates $\hat{Y} \in \mathbb{R}^m$

$$\widehat{Y}^{(3)} = f_\theta(X),$$

trained by empirical risk minimization using an absolute-error or squared-error loss. Note that this does not require that $f_{\theta,\pi}$ approximates Ψ_π, since it directly learns the mapping from signals to sleep metric.

8.2.4 Sleep staging models

We deployed five open source PSG sleep staging models on SHHS, all of which output hypnodensities as well as hypnograms from which sleep metrics can be calculated (estimation family 1). These are YASA, SleepLinear, SleepTransformer, and aSAGA. The models were trained to predict five sleep stage classes: Wake, N1, N2, N3, and REM, for 30-second epochs as per AASM recommendations. We trained and evaluated on in-distribution data from SHHS and used SRMS to benchmark models against human scorers, using a LOO human-consensus reference. In our implementation of the models and during training, we followed the description of the original methods [108], [110], [201], [270], [271]

All datasets were homogenized to match current AASM rules by merging Rechtschaffen & Kales sleep stage 4 into N3 in SHHS [8], [14]. Recordings were trimmed to within 30 minutes of the sleep period across all datasets, removing any reliance on lights-off and lights-on metadata. The models primarily use EEG signals (as well as EOG and EMG for some models). The available EEG derivations vary across datasets. In SHHS they are C3-A2 and C4-A1 and in SRMS F4-M1, F3-M2, C4-M1, C3-M2, O1-M2 and O2-M1.

The SleepLinear model [270] uses gradient-boosted decision trees (XGBoost), trained on engineered features rather than signal time series. The model uses EEG, EOG and EMG signals. Signals are resampled to 100 Hz and passed through a Chebyshev high-pass filter. Features are computed using time-domain statistic (e.g. interquartile range; IQR, skewness, kurtosis, Hjorth mobility/-complexity, Higuchi/Petrosian fractal dimension, and permutation enthropy) and frequency-domain bandpowers (sub-delta, delta, theta, alpha, sigma and beta) computed over multiple window sizes (30, 60 and 90 seconds), as well as ratios such as α/θ and δ/β.

In the SRMS dataset, where multiple EEG derivations are available, features were computed iteratively across all available EEG traces during training and evaluation. Training uses the standard multiclass XGBoost objective with standard tree regularization parameters, early stopping enabled, and cross-validation-based model selection.

TinySleep [271] is a deep learning model that uses a single EEG channel. It consists of a one-dimensional convolutional neural network (1D CNN) feature extractor followed by a long short-term memory network (LSTM) and a 5-class softmax classifier. Signals are resampled to 100 Hz and high-pass filtered using a Chebyshev filter. TinySleep is trained with sparse softmax cross-entropy, Adam optimization with gradient clipping, and dropout and L2 weight decay regularization. Model selection is done with periodic evaluation on a validation set and early stopping based on a combined criterion of accuracy and macro-F1.

SleepTransformer [108] is a hierarchical transformer model that operates on time-frequency representations at two levels, within-epoch spectrograms and across-epoch temporal structure. SleepTransformer uses EEG, EOG and EMG signals. Signals were resampled to 100 Hz and high-filtered using a Chebyshev filter. The EEG and EOG signals are bandpass filtered (0.3-40 Hz) and the EMG signal is high-pass filtered with a higher, 10 Hz cutoff. Each epoch is converted into a short-time Fourier transform (STFT) spectrogram using a fixed analysis window length L, hop size H, and fast Fourier transform (FFT) length n_{FFT}, producing a consistent time–frequency tensor with $n_{\mathrm{FFT}}/2 + 1$ frequency bins (resolution $\Delta f = f_s/n_{\mathrm{FFT}}$) per epoch. Invalid epochs are rejected if the spectrograms contain non-finite values. Training uses softmax cross-entropy averaged over sequence length with Adam optimization, dropout in the attention and dense blocks, and L2 regularization, with early stopping based on a validation set.

aSAGA [110] implements a U-Time fully convolutional encoder-decoder architecture for time series segmentations, with atrous spatial pyramid pooling to capture multi-scale temporal context. It uses the EEG signals, resampling to 64 Hz and filtering with the same Chebyshev highpass filter as the other models. The signals are normalized using $(x - \mathrm{median})/\mathrm{IQR}$ standardization. Training uses sparse categorical cross-entropy with Adam optimization and model selection is done with early stopping based on accuracy on

a validation set.

YASA [201] is an open source sleep staging pipeline which includes a gradient-boosted tree classifier (LightGBM) trained on engineered features from EEG, EOG and EMG signals. We used the included `yasa.SleepStaging().get_features()` method to compute the 149 features used and trained the LightGBM model for 400 boosting rounds with a learning rate of 0.05 and 63 leaves, in accordance with the YASA documentation.

8.2.5 Soft aggregation of hypnodensity

In the second sleep metric estimation family, estimation via hypnodensity, sleep metrics are computed from a staging model's epoch-by-epoch probabilities (hypnodensity), rather than a hypnogram. Hard aggregation of the hypnodensity collapses the probabilities to a hypnogram and computes sleep metrics as a point estimate. Soft aggregation instead treats the sleep stage at each epoch as a categorical random variable with probabilities $\hat{q}_t$ and propagates this epoch-by-epoch uncertainty through the sleep metric mapping Ψ_π, giving the sleep metric as a probability distribution. Soft aggregation thus provides a sleep metric uncertainty distribution of $Y_\pi|X$, which is not available under hard aggregation.

For additive sleep metrics (which are the focus of this chapter), it is in principle possible to compute the full distribution of $Y_\pi|X$ analytically. The total epoch count $\sum_t u_\pi(S_t)$ follows a Poisson-binomial distribution (assuming epoch-independence), whose probability mass function can be obtained exactly via probability generating functions, with $G_T(z) = \prod_{i=1}^{N}((1-p_i)+p_i z)$, where p_t is the epoch-by-epoch probability of belonging to the sleep stage or set of sleep stages in question [272].

In our implementation, however, we estimate the sleep-metric distribution using stochastic sampling from the hypnodensity (Figure 8.2). This approach has the advantage of being applicable to path-dependent sleep metrics, as well as additive ones. Here, we generate $B = 10,000$ sampled hypnograms by drawing $S_t^{(b)} \sim \text{Categorical}(\hat{q}_t)$ independently across all epochs t, and then computing the sleep metric on each sampled sequence, $Y_\pi^{(b)} = \Psi_\pi(S_{1:T}^{(b)})$. This gives us an empirical approximation to $P(Y_\pi|X)$ represented as a normalized histogram over the B samples.

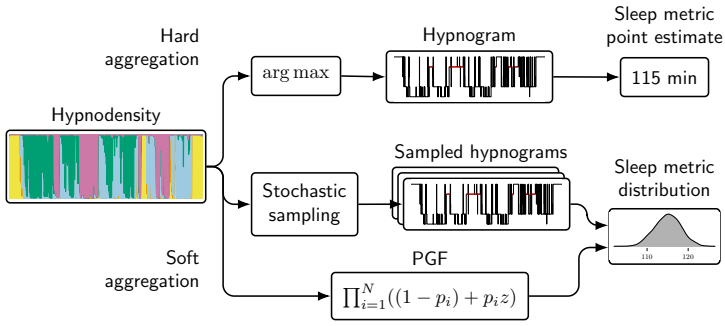


Figure 8.2: Soft and hard aggregation of hypnodensity for sleep metric estimation. A sleep staging model outputs a hypnodensity (epoch-by-epoch sleep stage probabilities). Hard aggregation first collapses this output to a hypnogram by arg max decoding and then computes a single point estimate of a sleep metric (e.g. total minutes in N3 sleep). In contrast, soft aggregation treats each epoch as a categorical random variable, with the probability of each sleep stage p_i, and propagates this uncertainty to the sleep metric level, yielding a distribution over the sleep metric. The metric distribution can be approximated by stochastic sampling of hypnograms from the hypnodensity or computed using probability generating functions (PGFs).

8.2.6 Direct estimation models

Heuristic signal intensity model

The heuristic signal intensity (HSI) model is a low-capacity, physiologically motivated estimator that maps a small set of handcrafted signal-intensity traces to sleep stage probability curves over time. Rather than learning an end-to-end mapping from raw PSG to epoch labels, HSI (1) extracts interpretable EEG/EOG/EMG traces, (2) maps these traces to sleep-stage-related probabilities using spline-based logistic models, and (3) estimates sleep metrics by integrating the resulting probability curves and applying a final linear calibration, as shown in Figure 8.3. Although we compute probabilities on 30 s windows for comparability with conventional staging, this discretization is an implementation choice. Since HSI outputs probability traces that can be integrated at any resolution, it is an example of an estimator in estimation family 3 (direct estimation of sleep metrics).

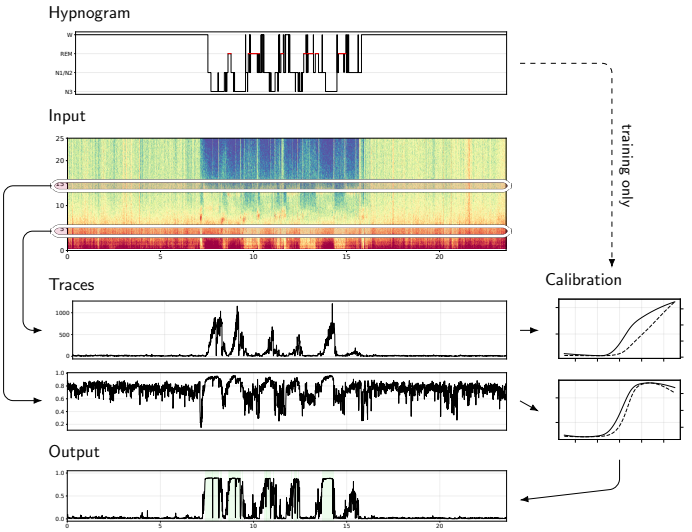


Figure 8.3: Heuristic signal intensity (HSI) model overview. From the raw input data (shown here in spectrogram representation), HSI extracts a small set of physiologically motivated signal intensity traces, by integrating the power within selected frequency bands (highlighted in pink). During training, the reference hypnogram (top) is used to fit calibration curves (right) that map trace values to sleep stage probabilities. At inference, the calibrated probabilities are combined to produce a sleep stage probability (bottom), shown here for N3 sleep stage. Other abbreviations: EEG, electroencephalography; EOG, electrooculography; EMG, electromyography; REM, rapid eye movement; W, Wake.

HSI uses three modalities: a single EEG channel, a single EOG channel, and a single EMG channel. Signals are loaded at their native sampling frequencies (no resampling). Trimming is optional; when enabled, recordings are trimmed to within 30 minutes of the sleep period (first non-wake epoch to last non-wake epoch, expanded by ± 30 minutes), consistent with the trimming used for the staging models. The signals are segmented into non-overlapping windows of length $\Delta t = 30$s. Prior to trace extraction, EEG and EOG are bandpass filtered using a 4th-order Butterworth filter (EEG: 0.3-35 Hz, EOG: 0.3-15 Hz). EMG is left unfiltered.

For each window t, EEG and EOG power spectral density (PSD)

estimates are computed using Welch's method (4 s Hann windows, 50% overlap; $n_{\mathrm{perseg}} = 4f_s$, $n_{\mathrm{overlap}} = 2f_s$; one-sided spectrum for real input), after mean-centering the window ($x_t \leftarrow x_t - \overline{x}_t$). Spectra are restricted to $0.5 \le f < 30$ Hz by selecting frequency bins (no interpolation). Band power for $[f_1, f_2)$ is obtained by numerically integrating the PSD over that band, and relative band power is formed by dividing by broadband power $B_t([0.5, 30))$ with stabilization constant $\varepsilon = 10^{-12}$:

$$R_t([f_1, f_2)) \;=\; \frac{B_t([f_1, f_2))}{B_t([0.5, 30)) + \varepsilon}.$$

For each window t, we compute a trace vector $\boldsymbol{\phi}_t \in \mathbb{R}^7$ with components $\phi_{t,1}, \ldots, \phi_{t,7}$:

- Delta-wave trace ($\phi_{t,1}$):

$$\phi_{t,1} = R_t^{(\mathrm{EEG})}([0.5, 4)).$$

- Theta low-amplitude mixed-frequency (LAMF) trace ($\phi_{t,2}$):

$$\phi_{t,2} = R_t^{(\mathrm{EEG})}([4, 8)).$$

- EOG-theta trace ($\phi_{t,3}$):

$$\phi_{t,3} = R_t^{(\mathrm{EOG})}([4, 8)).$$

- EEG spectral-slope trace ($\phi_{t,4}$):

$$\phi_{t,4} = \mathrm{slope}\Big(\log_{10}(f), \log_{10}\Big(\widehat{S}_t^{(\mathrm{EEG})}(f)\Big)\Big), \quad 2 \le f < 30 \text{ Hz}.$$

- EMG trace ($\phi_{t,5}$):

$$\phi_{t,5} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\Big(x_t^{(\mathrm{EMG})}[n]\Big)^2}.$$

- Broadband-energy trace ($\phi_{t,6}$):

$$\phi_{t,6} = \log_{10}\Big(10^{12} B_t^{(\mathrm{EEG})}([0.5, 30)) + \varepsilon\Big).$$

- High/medium-frequency ratio trace ($\phi_{t,7}$):

$$\phi_{t,7} = \log\left(\frac{B_t^{(\mathrm{EEG})}([25,30))}{B_t^{(\mathrm{EEG})}([10,15)) + \varepsilon}\right).$$

where $\log(\cdot)$ is the natural logarithm and $\log_{10}(\cdot)$ is base-10. EMG is not mean-centered prior to root mean square (RMS) calculation, so direct-current (DC) offsets contribute to $\phi_{t,5}$. Artifact masking is derived from the broadband-energy trace $\phi_{t,6}$ using robust per-night statistics. Let $\tilde{\mu}$ be the per-night median of $\phi_{t,6}$ and let MAD be the median absolute deviation; define $\tilde{\sigma} = 1.4826 \cdot \mathrm{MAD}$ (the usual MAD-to-standard deviation (SD) normal-consistency constant). A window is flagged as noisy with a noise mask $m_t^{\mathrm{noise}} \in \{0, 1\}$ if

$$m_t^{\mathrm{noise}} \;=\; \mathbf{1}\Big\{|\phi_{t,6} - \tilde{\mu}| > 3\tilde{\sigma}\Big\}.$$

In addition, a sustained-wake prior is computed from the ratio trace $\phi_{t,7}$ by applying a median filter (width 5 windows), thresholding using Otsu's method, and enforcing temporal persistence via binary opening with a structuring element corresponding to 60 minutes. For windows flagged by either mask, sleep stage probabilities are overridden to $\Pr(\mathrm{Wake}) = 1$ and $\Pr(\cdot) = 0$ for all sleep states.

To reduce between-subject and between-device variability, continuous traces were standardized within each recording using robust statistics computed on clean windows only (excluding windows where $m_t^{\mathrm{noise}} = 1$). For each trace dimension d, let median_d be the median and $\mathrm{IQR}_d = q_{0.75} - q_{0.25}$. We define a robust scale estimate

$$\hat{\sigma}_d \;=\; \max(\mathrm{IQR}_d/1.349,\, 1),$$

where 1.349 is the normal-consistency constant that converts interquartile range to a standard-deviation equivalent under a Gaussian assumption, and compute standardized, clipped traces

$$z_{t,d} \;=\; \mathrm{clip}\left(\frac{\phi_{t,d} - \mathrm{med}_d}{\hat{\sigma}_d},\, -5,\, 5\right).$$

The resulting transformation is applied to all windows (clean and masked), after which masked windows are overridden to Wake by the prior.

HSI maps standardized traces to sleep stage probability curves using separate ridge-regularized spline-logit models for three states: Wake, Deep (N3), and REM. Light sleep (N1/N2 pooled) is not modeled directly and is defined as residual probability mass.

Each standardized feature $z_{t,d}$ is expanded using a cubic B-spline basis with internal knots $\{-2, -1, 0, 1, 2\}$ and boundaries $[-5, 5]$. For state $s \in \{\text{wake}, \text{deep}, \text{REM}\}$,

$$p_t^{(s)} \;=\; \sigma(\eta_t^{(s)}), \qquad \eta_t^{(s)} \;=\; \sum_{d \in \mathcal{D}_s} \sum_{k} \beta_{d,k}^{(s)}\, b_k(z_{t,d}) + \beta_0^{(s)},$$

where $\sigma(\eta) = \frac{1}{1+e^{-\eta}}$, $b_k(\cdot)$ are spline basis functions, $\beta_{d,k}^{(s)}$ are their weights, and $\beta_0^{(s)}$ is an intercept. Feature sets are: Wake uses $\{\phi_{t,7}, \phi_{t,6}\}$; Deep uses $\{\phi_{t,1}, \phi_{t,4}\}$; REM uses $\{\phi_{t,2}, \phi_{t,3}, \phi_{t,5}\}$. Coefficients are fit by minimizing penalized negative log-likelihood with $\lambda = 10^{-2}$, applying L2 penalties to spline coefficients only (intercepts unpenalized), using limited-memory Broyden–Fletcher–Goldfarb–Shanno with bounds (L-BFGS-B) optimization.

Modeled probabilities are scaled by global multipliers a_s ("stage scales") which are defined as

$$\tilde{p}_t^{(s)} \;=\; a_s\, p_t^{(s)}, \qquad s \in \{\text{wake}, \text{deep}, \text{REM}\}.$$

If $\sum_s \tilde{p}_t^{(s)} > 1$, values are normalized proportionally. The remaining mass is assigned to Light sleep,

$$q_{t,\text{light}} \;=\; 1 - \big(q_{t,\text{wake}} + q_{t,\text{deep}} + q_{t,\text{REM}}\big),$$

yielding four probability curves $q_{t,\pi}$ that sum to one within each window. Windows flagged by the artifact/wake priors override $q_{t,\text{wake}} = 1$ and all others 0.

Sleep metrics are obtained by integrating each probability curve over time to compute raw minutes per state,

$$\widehat{Y}_\pi \;=\; \frac{\Delta t}{60} \sum_t q_{t,\pi},$$

and then applying a final linear calibration fit by ordinary least squares on a held-out SHHS validation set:

$$\widehat{Y}_\pi^{\,\text{cal}} \;=\; a_\pi + b_\pi\, \widehat{Y}_\pi, \qquad \widehat{Y}_\pi^{\,\text{cal}} \leftarrow \max(0, \widehat{Y}_\pi^{\,\text{cal}}).$$

Table 8.2: Approximate learnable parameter counts for the models used. Abbreviations: HSI, heuristic signal intensity; ViT, Vision Transformer.

Model Name	Number of Parameters
TinySleepNet	$\sim 3.0 \times 10^5$
SleepTransformer	$\sim 2.2 \times 10^7$
SleepLinear	$\sim 1.0 \times 10^5$
ViT	8.6×10^7
aSAGA	$\sim 1.5 \times 10^6$
YASA	$\sim 1.0 \times 10^5$
HSI	67

Optionally, the same probability curves can be assembled into a per-window distribution over states (a hypnodensity-like representation) and converted to a discrete hypnogram (e.g., by argmax) for visualization or compatibility with epoch-by-epoch analyses.

Table 8.2 highlights the large spread in model capacity across the compared methods. In particular, HSI has only 67 learnable parameters, several orders of magnitude fewer than the other models, which range from approximately 10^5 to 10^7 parameters.

Vision Transformer model

We implemented a high-capacity direct sleep estimation model (estimator family 3), based on a Vision Transformer (ViT) architecture [273]. The model was trained to predict sleep metrics from images of EEG and EOG spectrograms, without any epoch-by-epoch staging. The ViT splits the input images into fixed-size patches, embeds each patch as a token, adds positional encodings, and processes the token sequence using stacked Transformer encoder layers with multi-head self-attention and feed-forward blocks as shown in Figure 8.4 [273].

The model training is enhanced using the masked autoencoders method [274]. It introduces random masked patches of the input spectrogram, then the ViT is applied to encode the information, followed by a decoder that is attached only during training. The decoder task is to recompose the input, helping the encoder to learn from spectrogram spatial dependencies. By randomly sampling

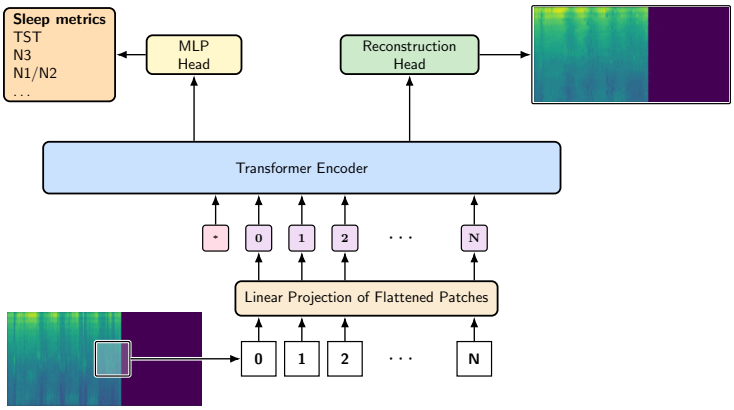


Figure 8.4: Vision Transformer (ViT) architecture used for direct estimation of sleep metrics from two EEG and one EOG spectrogram inputs, padded to 12 hours. The input spectrograms are partitioned into N fixed size patches, which are flattened and mapped to token embeddings via a linear projection. A learnable class token and positional embeddings are added and the resulting token sequence is passed to an MLP regression head to predict normalized sleep metrics, while a reconstruction head is used for self-supervised pretraining by reconstructing the spectrogram. This figure is adopted from [273]. Other abbreviations: EEG, electroencephalography; EOG, electrooculography; MLP, multilayer perceptron; TST, total sleep time; N1/N2, sleep stages N1 and N2; N3, sleep stage N3.

masks during the training, this method enforces a self-supervised training, augmenting the number of examples processed by the ViT encoder.

EEG and EOG signals are read from European Data Format (EDF) using the MNE Python library [204] and converted into time-frequency representations using code based on the spectrogram functionality in the YASA library. Spectrograms were computed using a multitaper spectrogram with 30 s windows, no overlap and frequencies restricted to 0.5-25 Hz, log-scaled to decibels (dB) and normalized per recording. Each spectrogram was zero-padded to a duration of 12 hours. During training, the placement of padding before and after the signal was randomized. The spectrograms are then resized to 224x224 pixels via interpolation. Non-finite values were removed and min-max normalization was applied to

each spectrogram to map all values to the range $[0, 1]$.

The ViT model was initialized with weights from ImageNet-21k pretraining. The pooled output 768 dimensional embedding is passed to a three-layer fully connected regression head, with progressively reduced dimensionality and rectified linear unit (ReLU) activation functions, to output sleep metric predictions. The model also has an optional reconstruction head used for self-supervised pretraining similar to a masked autoencoder. Here, the model learns to reconstruct masked input patches.

Sleep metrics were computed from the manually annotated sleep stages in the training data and normalized prior to training using standard scaling. The model was trained by minimizing mean squared error between the predicted and the reference sleep metrics. In the multitask (self-supervised) setting, the model has two branches (reconstruction branch and regression branch). In this setting, training minimizes a weighted combination of reconstruction loss and regression loss, controlled by a weight hyperparameter.

8.2.7 Experiments

The experiments performed use the two PSG datasets for different purposes. SHHS is used for model training and in-distribution evaluation, and SRMS for quantifying inter-scorer variability in sleep metrics and benchmarking models against individual human scorers under a leave-one-out consensus reference. All dataset splits are performed at the subject level. SHHS is divided into a training split and a held-out test split containing 20% of subjects. During the training of the sleep staging models and the ViT model, model selection is performed on the training split using K-fold cross validation. The HSI model is trained on 10 randomly selected subjects of the test set only and evaluated on the full test set.

Model robustness to dataset shift is assessed by applying the SHHS trained models to SRMS without retraining or fine-tuning, where recordings are collected using different sensors and annotated by a different human scorer.

To benchmark models' performance against individual human scorers, we evaluate on SRMS using a leave-one-out (LOO) consensus reference, where for each scorer sleep metric consensus is constructed from the remaining scorers and the scorer's MAE is compared the models' MAE relative to this reference.

We also perform a small-data SRMS experiment using the HSI model, where the HSI model parameters are fit using the first 5 records in the SRMS dataset on scorer 1's labels, and then evaluated using the remaining 45 records under the same LOO protocol, except scorer 1 is excluded entirely (i.e. scorers 2 - 10 are compared to the consensus of the remaining eight scorers).

Training of the ViT model and all staging models was conducted in PyTorch on the Icelandic High Performance Computing Centre (IHPC) infrastructure using a single NVIDIA A100 graphics processing unit (GPU) and Intel Xeon Gold 6248R central processing units (CPUs).

8.2.8 Evaluation and statistical analysis

Sleep metric model performance is reported as absolute error in minutes, summarized by mean absolute error (MAE) and the 95th percentile of absolute errors across recordings. In the SHHS data, one human scored the reference hypnogram per record, and sleep metric errors are computed relative to the sleep metrics derived from the human-scored hypnogram. For comparison, we show the empirical distribution of absolute pairwise disagreement between human scorers in the SRMS dataset as an external descriptive benchmark of human variability, using the metric-level pairwise comparisons from Chapter 7). On SRMS, errors are computed relative to the LOO reference sleep metrics, both on the models and the left-out human scorer. In some cases, hypnograms are compared using epoch-by-epoch staging accuracy and Cohen's κ. We report errors of a baseline naive constant predictor using the pooled healthy-control averages reported by Boulos *et al.* [247], implemented as described in Chapter 7.

To evaluate whether soft aggregation improves sleep metric estimation compared to hard aggregation, we performed a paired pooled comparison across all hypnogram-based sleep staging models (not including our ViT and HSI models). For each recording, model and sleep metric, we computed the paired difference in absolute error between the hard and soft aggregation and evaluated whether the difference was greater than zero using a one-sided Wilcoxon signed rank test. To calculate an approximate 95% confidence interval for the mean paired difference between the HSI model MAE and individual human scorers across the $n = 45$ evaluated records in SRMS, computed as a margin of

error MOE $= t_{0.975,44} \cdot \text{SEM}$ with the standard error of the mean calculated as $\text{SEM} = \text{sd}(d)/\sqrt{n}$ and $t_{0.975,44} \approx 2.01$, where d is the vector of record-wise paired differences (human absolute error minus model absolute error).

On SRMS, evaluation was performed relative to a leave-one-out metric-mean consensus reference. For each recording and each metric, we computed absolute error both for the model estimate and for the left-out human scorer relative to the corresponding leave-one-out reference. This allowed direct comparison of model-versus-consensus and scorer-versus-consensus performance on the same recordings. Sleep metric model performance was summarized using mean absolute error (MAE) in minutes and the 95th percentile of absolute error across recordings. On SHHS, where only one human-scored reference hypnogram was available per recording, errors were computed relative to the sleep metrics derived from that reference hypnogram.

8.3 Results

8.3.1 Preserving uncertainty improves sleep metric estimation

We evaluated whether preserving epoch-by-epoch uncertainty improved estimation of downstream sleep metrics. Five sleep staging models (YASA, SleepLinear, TinySleep, SleepTransformer and aSAGA) were retrained on SHHS and compared under two estimation strategies: conventional hard aggregation, in which the hypnodensity is collapsed to a hypnogram by argmax decoding before sleep metric calculation, and soft aggregation, in which sleep metrics are computed directly from the epoch-by-epoch probabilities.

The results of this comparison on the SHHS test set are shown in table 8.3. The retrained models achieved sleep stage classification performance as expected (accuracy = 0.78-0.86, κ = 0.67-0.79). Pooled across all models and sleep metrics, soft aggregation gave a statistically significant (although small) reduction in absolute error relative to hard aggregation (median improvement 0.03 min, mean improvement 0.52 min, p-value $= 3.6 \cdot 10^{-18}$). Soft aggregation had the most consistent positive effect on REM duration estimation, where it reduced errors across all models (median improvement

Table 8.3: Results on the SHHS dataset. Metrics reported are the mean and 95th percentile of absolute error of total sleep time (TST), and sleep-stage-wise errors (N1/N2, N3, REM). Abbreviations: ACC, accuracy; HSI, heuristic signal intensity; MAE, mean absolute error; P95, 95th percentile of absolute error; REM, rapid eye movement; SHHS, Sleep Heart Health Study; TST, total sleep time; ViT, Vision Transformer; κ, Cohen's kappa.

Method	**Staging**		**TST**		**N1/N2**		**N3**		**REM**	
	ACC	κ	MAE	P95	MAE	P95	MAE	P95	MAE	P95
YASA	–	–	21.0	93.1	29.6	78.1	19.5	67.9	15.0	44.0
via hypnogram	0.81	0.72	20.4	97.2	29.1	86.5	19.6	71.2	14.3	46.0
SleepLinear	–	–	20.7	68.4	31.2	93.5	12.5	40.7	24.3	66.5
via hypnogram	0.78	0.67	20.9	74.7	33.6	103.5	12.4	41.5	26.8	78.2
TinySleep	–	–	14.2	50.4	**21.2**	**63.8**	11.9	37.6	**12.2**	**38.7**
via hypnogram	0.86	**0.79**	14.6	53.0	21.7	67.5	12.5	38.2	13.3	43.5
Sleep Transf.	–	–	**12.9**	**42.1**	22.1	67.3	**11.6**	**36.7**	12.5	42.0
via hypnogram	**0.86**	0.79	12.9	42.5	22.1	67.2	11.6	36.7	12.6	42.0
aSAGA	–	–	22.1	65.8	26.1	73.0	20.7	55.5	23.5	69.0
via hypnogram	0.83	0.75	20.3	66.5	28.0	79.0	20.5	58.5	24.4	73.2
ViT	–	–	15.5	43.8	29.1	77.3	22.7	60.0	14.9	38.8
via hypnogram	0.71	0.57	26.7	66.0	34.1	90.7	24.2	63.2	15.3	39.2
HSI	–	–	33.3	82.2	–	–	35.0	73.3	25.1	61.2
via hypnogram	0.64	0.47	34.2	78.5	40.7	91.5	34.9	73.6	24.9	61.5

0.8–1.3 minutes per recording). The model least improved by soft aggregation was SleepTransformer (median difference close to zero), consistent with the finding that it already places near-unit probability on a single sleep stage in most epochs [108]. Figure 8.5 shows the distribution of absolute sleep metric error across the models. It also shows the empirical distribution of absolute pairwise human disagreement from SRMS as a reference.

8.3.2 Direct estimation of sleep metrics is competitive

Direct models trained to estimate sleep metrics from PSG achieved error levels comparable to strong staging-based systems on SHHS, and both direct models performed worse when their outputs were forced through a hypnogram intermediate.

In addition to soft aggregation of retrained staging models, we trained the ViT and HSI models which predict sleep metrics directly. On SHHS, both approaches achieved sleep-metric error levels comparable to the staging models (see table 8.3). On the SHHS test data, both models achieved sleep metric error levels comparable to the retrained staging models. When the two models estimated sleep metrics through a hypnogram intermediate (by first decoding epoch-by-epoch sleep stages and then computing sleep metrics), performance deteriorated.

8.3.3 Cross-dataset performance is limited

When evaluated on a separate multi-scorer cohort, none of the models trained on SHHS outperformed a naive constant predictor. Model performance on SHHS was evaluated against a single scorer per recording, so the resulting errors reflect both model limitations and scorer-specific variation in the reference labels. We therefore additionally evaluated all methods on the SRMS dataset, where multiple independent scorers were available for each recording. Following the metric-level reliability analysis in Chapter 7, we did not treat any individual scorer as ground truth. Instead, we used a leave-one-out (LOO) metric-mean consensus reference, defined for each comparison as the mean of the corresponding sleep metric across the other nine scorers. For each human scorer, we constructed a sleep metric consensus from the remaining nine scorers and computed the mean absolute error in sleep metrics.

We applied the SHHS-trained models to SRMS *as-is*, with no retraining or fine-tuning and evaluated against the LOO consensus reference of each human scorer. This way, model performance can be compared directly to human scorer performance. In this setting, most models failed to predict sleep metrics better than a naive constant predictor that outputs the global mean.

8.3.4 Direct sleep metric estimation under scorer disagreement

The SRMS dataset is too small to support training or retraining of the higher-capacity models used in this study, so for those methods it serves only as an external evaluation cohort. The HSI model is an exception because its parameterization is extremely small and can be fit from limited data. We therefore performed a separate small-data SRMS experiment in which HSI was fit using scorer 1's labels from the first five recordings and then evaluated on the remaining 45 recordings under the same leave-one-out consensus protocol, excluding scorer 1 from the reference. Despite this minimal training setup, HSI achieved errors close to the scorer-to-consensus range across the evaluated sleep metrics. Figure 8.6 shows a representative HSI output on SRMS, showing the nine-scorer majority-vote hypnogram along with the argmax-decoded HSI hypnogram and the corresponding probability traces for Wake, N3, REM and N1/N2, obtained by mapping physiologically motivated signal traces through calibration functions and normalizing.

Under this protocol, HSI achieved mean absolute errors of 30.8 min for TST, 35.7 min for N1/N2 duration, 23.1 min for N3 duration, and 19.6 min for REM duration. Compared with the scorer-to-consensus MAE values reported previously for SRMS (see Chapter 7), these errors were higher than the average human scorer for TST (14.23 min), N1/N2 duration (23.92 min), and REM duration (9.28 min), but were very close for N3 duration (22.71 min). Expressed relative to the observed range across individual scorers, HSI fell outside the scorer range for TST, N1/N2, and REM, but within the scorer range for N3. However, HSI still outperformed the naive constant benchmark for all four metrics (56.62, 45.53, 34.45, and 23.85 min, respectively). Figure 8.7 shows the MAE in each sleep metric for each of the nine scorers, when compared to the consensus of the remaining eight scorers.

Under this protocol, HSI's errors come close to the scorer-to-consensus disagreement of the human scorers for all sleep metrics. In a paired comparison against each scorer (using the same LOO consensus reference), HSI achieved lower MAE than two scorers in N3 duration prediction.

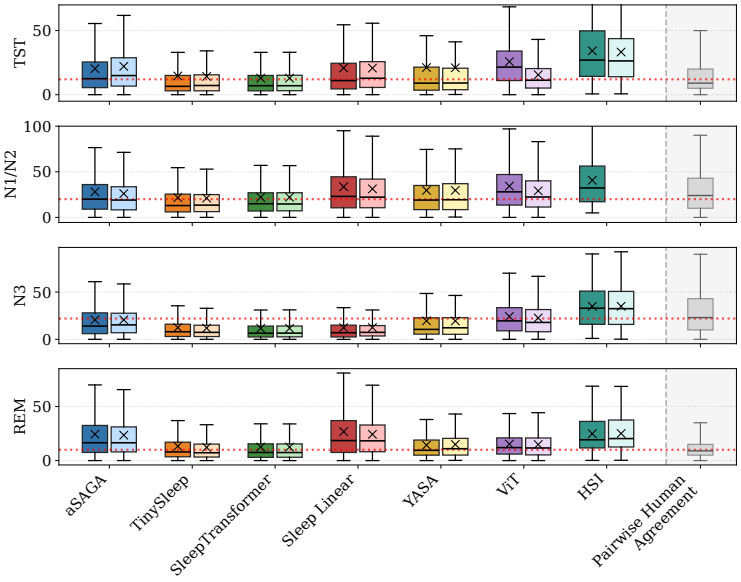


Figure 8.5: Sleep metric estimation error results on the test set of the Sleep Heart Health Study (SHHS) dataset. The boxplots show the absolute error distribution for hard aggregation (darker colored boxes) and soft aggregation (lighter colored boxes). Each panel shows the absolute error in minutes for total sleep time (TST) and sleep stage durations (N1/N2, N3, and rapid eye movement [REM]) across all trained sleep staging models. The rightmost gray boxplot shows the distribution of absolute pairwise human agreement in the Sleep Revolution Multicenter Scoring (SRMS) dataset, included as an external human-variability benchmark derived from the metric-level reliability analysis in Chapter 7; the red dotted line marks its mean.

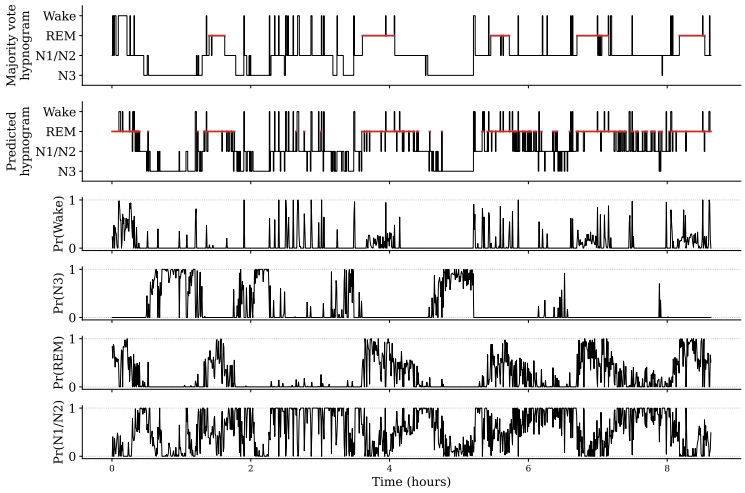


Figure 8.6: Example output of the heuristic signal intensity (HSI) model on a single recording from the Sleep Revolution Multicenter Scoring (SRMS) dataset, illustrating how HSI produces a probabilistic representation of sleep architecture. The top panel shows a 9-scorer majority-vote hypnogram used as reference, and the second panel shows the HSI-derived hypnogram obtained by argmax decoding of HSI's predicted sleep stage probabilities. The lower panels show HSI's posterior probability traces for Wake, N3, REM, and N1/N2 over the recording. These probabilities are computed by mapping a small set of physiologically motivated signal-intensity traces (e.g., band-power and ocular/muscle activity surrogates) through learned calibration curves and normalizing to a valid hypnodensity (probabilities sum to 1). Other abbreviations: REM, rapid eye movement; N1/N2, sleep stages N1 and N2; N3, sleep stage N3.

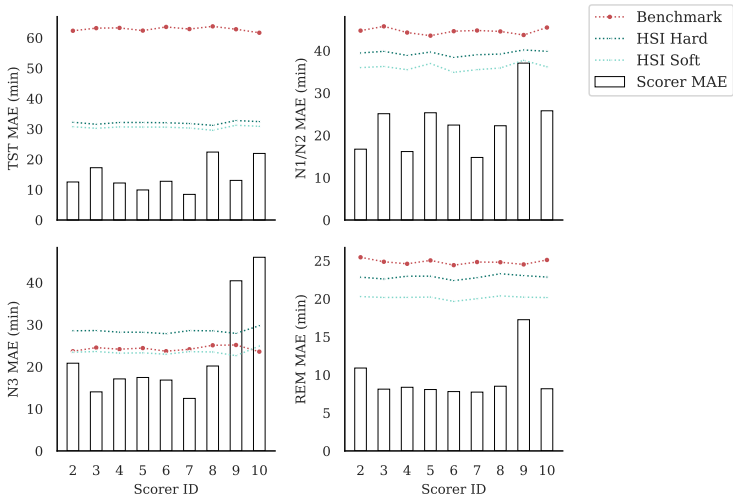


Figure 8.7: Leave-one-out comparison of sleep metric estimates against a human consensus reference in the Sleep Revolution Multicenter Scoring (SRMS) dataset, excluding scorer 1. For each scorer ID, the consensus of the other eight scorers is treated as reference. The bars show the scorer's mean absolute error (MAE) in minutes in sleep metrics relative to this reference. Dotted green lines show the corresponding errors for the heuristic signal intensity (HSI) model (soft and hard aggregation). The dotted red line shows the error of a naive constant predictor (global mean baseline). HSI was trained on scorer 1's scorings of five records from this dataset and is evaluated on the remaining 45 nights. The four panels show MAE for total sleep time (TST) and sleep stage durations (N1/N2, N3, and rapid eye movement [REM]).

8.4 Discussion

The main finding of this study is that when the primary objective is to determine a sleep metric, treating the hypnogram as an intermediate representation can introduce avoidable error. Across the experiments in this work, methods that reduced reliance on hard sleep-stage-label decoding either improved sleep metric estimation or remained competitive with conventional staging-based pipelines. Taken together, these results support the view that sleep metric estimation should be treated as a primary prediction task in automated sleep analysis, rather than only as a downstream consequence of sleep staging (see Chapters 6 and 7).

The clearest evidence for this comes from comparing hard and soft aggregation of hypnodensity. In that setting, the upstream staging models, training data, and model outputs were unchanged; only the mapping from probabilistic sleep stage output to sleep metrics differed. Soft aggregation, therefore, isolates the effect of the intermediate representation itself. Across models and metrics, propagating probabilistic sleep stage information reduced absolute error relative to hard hypnogram decoding, with the clearest and most consistent gains observed for REM duration. This suggests that some of the errors in estimated sleep metrics do not arise solely from imperfect upstream staging, but also from collapsing uncertain sleep stage predictions into a single label sequence before the metric is computed [111], [195], [196], [255].

The direct estimation experiments on SHHS support the same general conclusion from a different angle. Both the ViT model and the HSI model achieved sleep-metric errors within the same broad range as strong staging-based models when evaluated directly at the metric level. Moreover, when these direct models were forced back through a hypnogram-like intermediate representation, their performance deteriorated. This further supports the idea that optimizing for sleep metric estimation is not equivalent to optimizing for sleep stage classification and then deriving the metrics afterwards. For sleep metrics, the hypnogram is not always the most useful intermediate target (see Chapters 6 and 7) [256], [257].

The HSI and ViT results are notable for a second reason. Unlike the higher-capacity models, HSI has extremely low model capacity and is built around a small number of physiologically motivated signal traces. Its competitive performance on SHHS

suggests that at least some commonly used sleep metrics may be estimable from relatively simple signal summaries, without requiring a full high-capacity staging pipeline. At the same time, the ViT result is notable because it operated on only three 224×224 spectrogram images summarizing the full-night recording, rather than on the original high-frequency time-series signals across hundreds of 30-second epochs. Despite this compression of the input data, the model still achieved competitive sleep-metric estimation, in the same broad range as pairwise human disagreement, as illustrated by the gray boxplots in Figure 8.5. For example, the mean absolute errors for TST, N1/N2 duration, N3 duration, and REM duration were 15.5, 29.1, 22.7, and 14.9 min, compared with mean pairwise scorer differences of 18.56, 31.96, 30.61, and 12.30 min, respectively. This comparison should be interpreted cautiously, however, because the human disagreement values come from pairwise scorer differences on a different dataset and therefore serve only as a rough contextual benchmark rather than a direct reference standard. These results suggest that direct estimation can remain viable even when the model is trained on a highly condensed representation of the overnight PSG. We do not interpret this as evidence that simple models or condensed data representations are always preferable, but it does suggest that direct estimation of sleep metrics can be viable even with limited models or data [262]–[264].

At the same time, the results on the multicenter dataset show that bypassing the hypnogram does not solve the generalization problem. When models trained on SHHS were applied directly to SRMS without retraining, performance deteriorated substantially, and most methods did not outperform a naive constant predictor. This indicates that good in-distribution performance on a large development cohort does not automatically translate into robust performance on a new cohort [275].

A central issue in interpreting these results is that the target quantity itself is uncertain when derived from human scoring. Prior work has shown that hypnogram-derived sleep metrics vary meaningfully across scorers, and for this reason, we do not treat any single scorer-derived value as ground truth. Instead, on SRMS, we used a leave-one-out consensus defined at the sleep-metric level. This provides the closest observable proxy to the quantity of interest and allows model-versus-consensus and scorer-versus-consensus errors to be compared on the same scale. Under this

benchmark, the results can be interpreted relative to the level of disagreement already present among expert scorers, rather than relative to an arbitrarily chosen individual scorer (see Chapter 7) [17], [194].

This issue is not unique to sleep scoring. More broadly, work in medical informatics has argued that clinical data should not always be treated as fixed and uniquely determined ground truth, but rather as the outcome of recording practices, interpretation, and context-dependent decision processes [276]. From this perspective, disagreement among scorers is not merely annotation noise to be eliminated, but part of the reliability problem that any clinically meaningful artificial intelligence (AI) system must confront. Our use of leave-one-out metric-level consensus is consistent with this view: rather than privileging a single recorded value, it treats the target as an uncertain quantity that must be estimated relative to expert disagreement.

Due to its very small parameterization, the HSI model could be fit using scorer 1's labels from only five recordings and then evaluated on the remaining recordings under the same leave-one-out consensus protocol, excluding scorer 1 from the reference set. Even under this highly constrained setting, its errors were close to the scorer-to-consensus range across the evaluated sleep metrics. This result suggests that direct estimation of sleep metrics remains feasible even with limited labeled data.

More broadly, these findings raise the possibility that some useful sleep endpoints may be better modeled directly from physiological signals than indirectly through discrete sleep stage labels [258]–[260], [264]. The conventional staging pipeline and direct sleep metric estimation should be viewed as alternative routes, with the most appropriate choice depending on the intended endpoint.

These considerations are also relevant for recent transformer-based foundation models for sleep analysis, such as SleepGPT and SleepFM [185], [277]. Despite substantial increases in model scale, complexity, and training data relative to earlier automatic sleep-staging systems, reported performance on conventional sleep-staging benchmarks still tends to lie within a relatively narrow range. This suggests that sleep-stage-classification accuracy alone may be an insensitive benchmark for comparing models, particularly when the reference labels themselves are subject to substantial expert disagreement. At the same time, such foundation models appear to learn representations that transfer across multiple

downstream tasks, including sleep staging, spindle detection, and disease-related classification. From this perspective, the metric-level targets proposed in the present work may provide a useful additional way to train and evaluate such models, by testing whether improved representations translate into more reliable estimation of clinically relevant sleep endpoints rather than only marginal gains in sleep-stage-label agreement.

This study has several limitations. First, the experiments were conducted on a limited set of cohorts, and broader validation across additional populations, recording environments, and scoring settings is needed [275]. Second, the analysis focused primarily on additive duration metrics rather than path-dependent measures such as sleep onset latency or REM onset latency, which may behave differently under direct estimation or uncertainty propagation. Third, the poor transfer from SHHS to SRMS shows that model generalizability remains an open problem. More generally, the present results reinforce that performance estimates in sleep AI are constrained not only by model architecture and dataset size, but also by the reliability and clinical relevance of the target representation itself [276]. Finally, any proposed alternatives to conventional sleep metrics should ultimately be judged by their clinical usefulness and their ability to predict relevant outcomes, rather than only by agreement with the prevailing scoring protocol [259], [260], [264].

Future work should therefore proceed in two directions. One is methodological: extending direct, uncertainty-aware estimation of sleep metrics to a broader set of clinically used measures and validating these estimates across more heterogeneous cohorts [195], [196], [255]. The other is clinical: testing whether direct signal-based estimates of sleep metrics or alternative signal-derived sleep descriptors provide stronger or more reliable associations with downstream outcomes than conventional hypnogram-derived measures [258]–[260], [264].

8.5 Conclusion

When the endpoint of interest is a sleep metric, the hypnogram should not automatically be treated as the primary prediction target. In this study, approaches that reduced reliance on hypnogram decoding improved or matched conventional staging-based

pipelines at the metric level, showing that performance in sleep staging does not necessarily translate into performance in the estimation of downstream sleep metrics. At the same time, poor transfer to the external multicenter dataset showed that bypassing the hypnogram does not, by itself, solve the problem of generalizability. Automated sleep analysis should be developed and evaluated at the level of the clinical or research endpoint of interest, using reference standards that account for expert disagreement rather than assuming that a single scorer-derived value is ground truth.

Chapter 9

Human-AI collaboration

From explainable AI to co-creating meaning

Full reference: E. Hardarson *et al.*, 2024, "Human-AI Collaboration: From Explainable AI to Co-Creating Meaning", *ACIS 2024 Proceedings*

9.1 Introduction

Recent advances in machine learning (ML) have given rise to new forms of digital work [278], [279]. Such digital work can involve dynamic configurations of people and algorithms that collaborate in conducting tasks within practices across domains [280], also referred to as human-AI collaboration. As a result, there is a growing stream of literature on human-AI collaboration that explores the complementarities between people and algorithms in performing work and its challenges, such as human-AI team dynamics [281], task division between human and algorithmic agents [282], and establishing trust in human-AI relationships [283].

Much existing research assumes that human-AI collaboration can be improved by refining the explainability of ML algorithms to people in digital work configurations [284], such as by improving the transparency of ML decision making [283]. In essence, this

approach seeks to tackle a root problem of ML, which is that the process of generating these outputs is inscrutable [285] because they are inherently opaque and difficult to interpret [286]. As phrased by Faraj et al. (2018, p. 63): "*even if a learning algorithm is coded in a very simple and comprehensible way, understanding how it arrived at its result can still be unclear, even to its developers, due to the quantity and complex interaction of the data fed into the algorithm*". This problem is also known as "the black box" of AI [287].

However, this approach to explainable AI (XAI) has certain limitations for human-AI collaboration, which arise from the fact that ML processes are fundamentally different from human intelligence [288], [289]. The traditional approach to XAI typically addresses this problem by mimicking a domain-specific language in the creation of explanations of algorithmic processes to be persuasive to humans working in a specific context [290]. However, this is problematic because this approach risks producing an explanation that is untrue to how the ML output was generated since the process that underlies these algorithms is a statistical model that learns to recognize complex patterns in large volumes of data [129]. Against that backdrop, our aim is to outline an alternative approach to XAI that can enhance human-AI collaboration.

To do so, we draw on framing theory, a theory of meaning-making [291], to problematize a dominant approach to XAI in human-AI collaboration and outline an alternative approach. While the former rests on a persuasion-based approach, akin to strategic framing, the new approach rests on a co-creation approach in which humans and AI create meaning together, similar to interactive framing[292]. We selected framing theory because of the theory's past usefulness and future potential to approach questions of digital transformation in organizations [293]. In the chapter that follows, we outline and motivate this new approach by unpacking and contrasting the different frames that underlie (1) the practice of developing XAI that becomes encoded in ML algorithms by using existing literature on XAI as a proxy for this practice and (2) an existing practice of sleep diagnosis, which constitutes an empirical case of a domain practice in this research. The alternative co-creation approach to XAI constitutes the main contribution of this chapter.

9.2 Explainability in AI literature

XAI is a field of research that seeks to make black-box AI systems more transparent and understandable to humans. The need for XAI arises from the fact that AI systems, especially those based on deep learning (DL) algorithms, are inherently opaque and difficult to interpret [130]. DL algorithms and other ML models are not hard-coded rule-based systems but rather statistical models that learn to recognize complex patterns in training data [129]. Whereas traditional rule-based algorithmic systems are inherently explainable (illustrated in Figure 9.1), black-box ML models require additional explanation methods to be understandable to humans (illustrated in Figure 9.2).

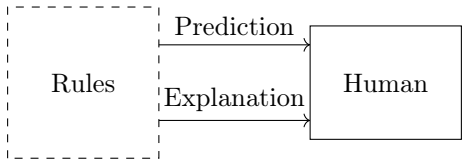


Figure 9.1: Transparent, rule-based model. Inherently explainable.

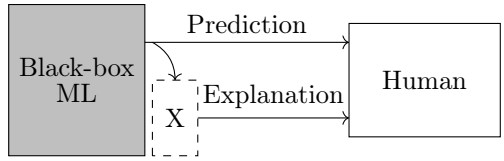


Figure 9.2: Black-box machine learning model. Explanation as strategic framing, based on convincing message created with explanation generating module, X.

There is no consensus in the literature on the definition of "explanation" in the context of XAI [133]. The "explainability" concept is itself ambiguous and has been subject to interpretation and philosophical discussion for centuries [134], [135]. Generally speaking, XAI aims to make AI systems more trustworthy by providing explanations [136]. The approaches to XAI can be divided into two main categories: interpretability and post-hoc explainability [130], [137]. Interpretability refers to how easily the model's internal decision-making process can be understood by a human and is an inherent property of the model. For example,

a simple linear regression model is interpretable. Conversely, post-hoc explainability refers to the methodology of generating explanations after the model has made a prediction [130]. In this chapter, we focus on text-based post-hoc explainability.

Various authors have criticized the post-hoc explainability approach to XAI [294]. Babic and Cohen [139] argue that post-hoc explanations are "insincere-by-design", unable to make transparent the actual reasoning behind an AI model's decision. Rudin [138] further criticizes post-hoc explanations for being unreliable and misleading and argues for the use of interpretable models instead. Nevertheless, research on post-hoc explainability continues to be one of the dominant approaches in XAI research [130].

The transparency of the model internals is a central topic of XAI research. According to Hulsen [136], the explanation should reflect the models' internals in a transparent way. The explanations given by an XAI system should truthfully explain the inner process of the AI model decision process [295], [296].

In XAI, explanations are viewed as an additional output of the AI system [139]. The literature describes XAI as an AI that *produces* reasons for its decision [296] or *provides* explanations to human users [130]. In this way, explanations are framed as one-way messages in the XAI literature. XAI systems are designed to convince the user to trust the system [130], [133]. Explanations are framed in a way that aligns with the background of different stakeholders [136]. As such, the explanations are designed to be persuasive and convincing. Merry *et al.* [297] point out that the language used in XAI explanations should take the audience's background into account and could be either generic or domain-specific. Jin *et al.* [295] argue that in a medical context, XAI explanations should use terminology that is clinically relevant. However, little research has scrutinized the implications of these assumptions of XAI for human-AI collaboration in practice. To do so, we now introduce framing theory as a lens to unpack XAI in this context.

9.3 A framing perspective on XAI

Framing theory is a theory of meaning-making [291]. It focuses on the application and formation of frames by different actors, including individuals, groups, organizations, and even fields (ibid.).

Frames have been described as interpretative schemes [298] and meaning systems [299], as they consist of and are structured by shared beliefs, assumptions, expectations, and pieces of knowledge [300]. They provide meaning to action, as they function as vehicles for interpreting experiences and guiding perceptions of action possibilities [301]. Frames also concern actors' views of specific technologies, referred to as technological frames [300].

Framing refers to social processes that shape how frames are structured and what they contain [302]. The literature has recognized different types of framing processes. Two of these are strategic and interactive framing [292], which we find useful for approaching human-AI collaboration. Strategic framing broadly centers on how one actor seeks to impact another actor's frame through persuasion. As such, strategic framing focuses on the design of unidirectional messages with rhetoric and language that aim to convince someone else to adopt a particular frame [303]. In contrast, interactive framing centers on meaning-making as a co-creation process, in which interactions between different actors' power framing [304]. As such, it focuses on the generative potential of exploring differences between frames via multiway interactions.

Additionally, people's interactions with technology play a central role in framing processes [299]. This is because actors' frames become inscribed in technologies via design choices [279], such as in choices of what elements to include in interface design and what data to use for training ML algorithms [305]. Thereby, frames become embedded in and diffused via the design of XAI interfaces. Different XAI interfaces can thereby reflect different views about what constitutes a proper explanation as well as contain different assumptions about what role the human actor working with the algorithm should play in the human-AI collaboration. While some XAI interfaces may be designed to persuade the actor of an algorithm's conclusion, as in conventional approaches to XAI (Figure 9.2), other interfaces may be designed to invoke co-creation of meaning together with the actor on how to interpret and respond to what the algorithm suggests [306]. One possible way to do so is that the interface signals ambiguities to the human, which triggers a cycle of reflection amongst the human who can, for example, complement the ML decision-making process with their unique intelligence such as human judgment [288] (Figure 9.3). This is the central idea we put forth in this chapter. We now proceed to demonstrate the need for an alternative approach with

an empirical study of a sleep scoring practice in sleep medicine, which reveals frame differences of XAI that complicate the reliance on traditional approaches to XAI.

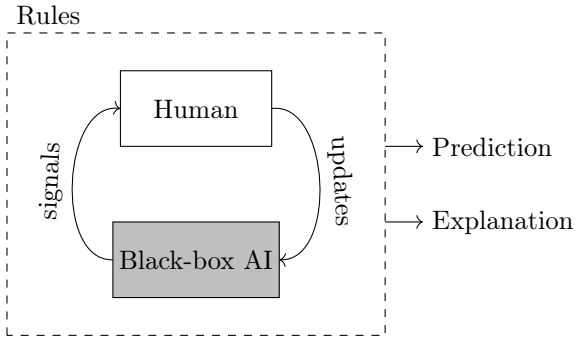


Figure 9.3: Co-creating meaning through human-AI collaboration within a rule-based framework.

9.4 Research method

9.4.1 Research context

In order to explore the framing of XAI in human-AI collaboration in a practical setting, we looked at the use of XAI in sleep medicine. In sleep medicine, like other highly regulated practices, clinicians must follow an established framework of rules and guidelines in their work diagnosing and treating sleep disorders. In this practice, the American Academy of Sleep Medicine (AASM) publishes rules for how sleep data should be analyzed for clinical diagnosis, including how sleep stages should be classified [209]. Sleep staging is the process of classifying 30-second segments of sleep recordings (epochs) into different sleep stages, such as light sleep, deep sleep and rapid eye movement (REM) sleep [209]. This process is an important step in the diagnosis of sleep disorders, such as sleep apnea and insomnia. It is usually performed by sleep technologists, who classify sleep epochs using visual inspection of various biosignals recorded during sleep, such as electroencephalography (EEG), electromyography (EMG) and electrooculography (EOG). The process of analyzing one night's sleep takes about two hours

for an experienced sleep technologist [12]. Although accurate automatic sleep staging algorithms have existed for several years, they are seldom used in clinical practice without thorough manual review. This is in part because of their black-box nature [15], [19]. Most of the automatic sleep staging methods use DL, and few (if any) are based on the AASM scoring rules [9], [199]. Sleep staging thereby represents a case of a domain practice [307] that experiences problems of explainability in emergent configurations of human-AI collaboration.

9.4.2 Data collection and analysis

Against this backdrop, we conducted a qualitative study with experienced sleep technologists to explore their views on XAI for sleep staging. An open call for participation in the study was sent to sleep experts involved in the Sleep Revolution project [160], specifically targeting individuals whose work involves the classification of sleep stages using EEG signals. Eight sleep experts participated in semi-structured interviews [308]. The participants had between 10 months and 20 years of experience in sleep staging (on average 10 years), working in Greece, Belgium, Finland, Sweden, and Iceland. Five of the participants were men and three were women.

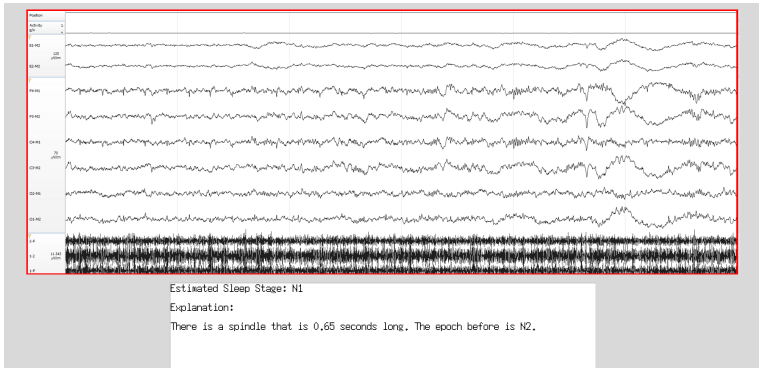


Figure 9.4: A screenshot of the interface shown during the interview. The participant sees the polysomnography output from the Noxturnal software and below it an estimated sleep stage with a synthetic explanation provided for interview discussion.

The interviews had three phases: *Ideation*, *exploration* and

reflection. During the ideation phase, the participants were asked about their background in sleep medicine and their familiarity with AI. They were asked to imagine using a hypothetical XAI sleep staging system in a clinical setting and share their thoughts about the explanations such an XAI system should offer. In the exploration phase of the interview, the participants were shown a mock-up graphical user interface for a hypothetical XAI sleep staging system (see Figure 9.4). The interface showed multiple 30-second segments of sleep data from an existing sleep analysis software called Noxturnal, (Nox Medical, Reykjavik, Iceland), along with an estimated sleep stage. The participants were also shown a hypothetical explanation, which was written by the researchers (see examples in Table 9.1). Ten examples were explored and each time the participants were asked if they agreed with the sleep stage classification, how clear and valuable they found the explanation, and whether the explanation matched their expectations. In this way, these hypothetical explanations were a way to make the participants reflect on and talk about 'what might be' in the scenario of integrating XAI in the clinical practice [309]. Finally, in the reflection phase, participants were asked general questions about XAI for sleep staging and open discussion was encouraged.

When creating the example explanations, we aimed to emulate different types of XAI explanations, such as attention-based, feature importance-based, and rule-based explanations, using the XAI taxonomies from Barredo Arrieta *et al.* [130] and Speith [310] as a guide. The explanations were all text-based and written to be plausible, but not necessarily correct. Table 9.1 shows some examples of the explanations used. Technical terminology for feature names such as "permutation energy" and "mean curve length" is adopted from Najdi *et al.* [311].

The interviews, which lasted around 45 minutes each, were performed online, recorded and transcribed verbatim. The transcripts, totaling 96 pages, were then analyzed in the Atlas.ti software using thematic analysis [312]. We used an abductive approach for the data analysis [313]. We reviewed the interview transcripts and used inductive coding to find themes relevant to how the participants viewed XAI for sleep staging. We then examined these thematic codes in the context of existing XAI literature. We used the notion of frames to guide our coding process, focusing on the views, beliefs, and expectations that guide the sleep experts in

Table 9.1: Examples of the explanations shown in the hypothetical XAI interface used in the interviews.

Example	**Type of Explanation**
The permutation energy of the C4-M1 and F4-M1 signals are above baseline and the relative power in the alpha band of the F4-M1 signal is below baseline. The algorithm has been trained to recognize these patterns as indicative of N2 sleep.	Interpretable model
35.0% of the epoch consists of slow wave activity in the F4-M1 and F3-M2 signals (between 4.9 and 13.7 seconds and between 22.3 and 24.0 seconds).	AASM rule-based
Classified as N2 with 70% confidence. Other suggested states were wake (21%) and N3 (9%).	Uncertainty quantification
The most important features were the Teager energy and the mean curve length of the last 15 seconds of the E1-M2 and E2-M2 signals, as well as the total power in the 2-F, 1-2 and 1-F signals.	Feature importance
Looking at the spectrograms of F4-M1 and F3-M2, the algorithm paid attention to the left side of the signal, between 12 and 15 Hz.	Attention
The algorithm used is a convolutional neural network (CNN) that was trained on a large dataset of PSG data.	Mechanistic

their work and shape their perspectives on XAI. Through this iterative process, we identified four distinct elements of these frames, which we elaborate on in the next chapter.

9.5 Explainability in clinical practice

The participants were all optimistic about the use of AI for sleep staging. Some were hopeful that AI would reduce the time spent on manual sleep staging, while others believed that AI could make the process more consistent and accurate. However, they also expressed concerns about the black-box nature of AI algorithms and spoke of the need for XAI. The participants also universally agreed with the need for AI systems to not just be accurate but also explainable. As Participant 1 summarized: "*I think [sleep staging AI] is definitely the way of the future. I think it is going to be growing with leaps and bounds. And I think it is going to streamline the way that we are doing things. It is something that can be very uncomfortable for those who have not used it in the past [... but ...] is definitely going to mitigate a lot of confusion and discrepancies [...]. So, if the AI is able to definitively give an explanation as to why it evaluates the stage a certain way, I think it will definitely lead to the acceptance of how that is scored*".

We will now detail the sleep experts' frames that guide their work of sleep staging across four dimensions, which shape their views on the possible application of XAI in this practice

9.5.1 Rule-based models or pattern recognition

Before interacting with the XAI interface, the participants were asked what type of explanation they would expect from an XAI sleep staging system. All participants except one expected the explanations to be based on the AASM scoring rules. For example, Participant 5 imagined such a system would have "*been trained to tick all the boxes from the guidelines from the AASM [...] and just use this grid like a list of checkboxes*".

Participants were often surprised when the first explanation they were presented with did not follow the AASM rules and terminology. The belief that AI sleep staging would operate using rule-based models following AASM rules was especially strong amongst some participants. When asked if AI systems that do not follow AASM rules should be avoided, participant 5 said, "*That can't be true because the model is trained based on the rules*".

However, the participants also recognized that sleep staging could, in some situations, be performed with pattern recognition. Indeed, the participants recognized that in some cases, they themselves and their colleagues use pattern recognition to classify sleep stages. For example, Participant 3 said scoring is based on a "*gut reaction*". According to Participant 5, "*we see the epochs and actually, we are trained to see patterns. It's pattern recognition [...] for us because we just click, click, click, click. [...] We have been trained, read books, and thousands of studies and we just see the patterns*". However, when asked to explain the reasoning behind assigning a sleep stage, all participants quoted AASM rules or similar.

9.5.2 Interactivity and multi-way communication

Some participants hinted towards a hybrid approach, where AI system would use ML pattern recognition in parallel with AASM rule-based algorithms. Participant 6 was asked if sleep staging AI based on pattern recognition should be avoided: "*I think that there should be maybe a middle way. Obviously if you train an algorithm*

with large datasets for pattern recognition, it is valuable and it can be used. But at the same time, I think there should be some sort of rule-based evaluation". This points towards an AI system where black-box ML models and humans collaborate within a rule-based framework. The participants also expressed a desire to ask further questions of the explanation, such as requesting more detailed information about the specific features used.

9.5.3 Communication via familiar language

Participants universally agreed that the explanation should use familiar terminology, grounded in the scoring rules. Participant 2 described the need for "*common language, [...] the terminology and not necessarily all the rules, but to use terms and words that I am familiar with*". This raises a central question: Should the explanation use the same terminology as the AASM rules, even if the AI system does not operate using the concepts and logic of the rules? Should the XAI explanation, for example, reference "sleep spindles" (a term from scoring rules describing a specific pattern in EEG signals) if the AI system does not use this phenomenon to classify sleep stages?

9.5.4 Transparency of model internals

The majority of participants did not describe a need for the explanation to reflect the AI system's internal logic. Instead, they preferred the explanation to "*give you an explanation that you are used to [...] even though that's not necessarily how the model is making its decision*" (Participant 1). Participant 2 said that it does not matter whether the explanation given by the AI system is the "truth", a true description of how the algorithm actually works, elaborating that "*it could use all these algorithms and digital tools to work on the decision, but I would like the explanation to be more familiar*".

The participants seemed to make a distinction between the method used by the AI system and the explanation provided by the used: "*AI is not a digital human brain. It uses completely different ways of calculating things. So I know that the method is this new thing, this innovation. So I would be ready to trust the method as long as the explanation was clear to me, I don't know. For example, I use my eyes to find a K-complex. I don't care which*

way the AI uses to find the K-complex as long as it tells me that this is a K-complex" (Participant 2).

Not all participants agreed with this view. According to Participant 8, the explanation should not make the user think the AI system is using features or phenomena from AASM scoring rules: "*if that is not actually what the algorithm is looking for or what the algorithm's decision is based on, I think it's kind of lying*". Participant 6 spoke in a similar vein of the danger of the AI explanation being misleading: "*You shouldn't make the AI explain something if it's not really using [it]. For example, it maybe shouldn't be trying to force out an explanation based on the rules because [...] it could start just lying to you, making stuff up. That could be a danger or problem. [...] It could be dangerous to force it to come up with explanations if it starts saying things that really were not correct*".

9.6 Discussion

As the prevalence of human-AI collaboration expands across practices, located in many different sectors and industries, the need to improve knowledge on topics such as human-AI team dynamics [281], human-AI task division [282], and human-AI trust [283] will grow. While XAI has emerged and solidified as a general solution to improving human-AI collaboration, we still know surprisingly little of how it can be integrated into and enhance practices within different work contexts, each shaped by unique contextual factors such as rules, norms, and cultures. Such contextual differences can also be observed in the process of sleep scoring across various countries or even workplaces [15], [16]. Therefore, we saw the need for taking a critical look at existing, dominant assumptions of XAI and investigating responses to it amongst practitioners within a specific domain context, namely sleep medicine. The challenges of dominant approaches to XAI also became apparent in the interviews where the participants expressed skepticism to how explanations are typically constructed in XAI. These findings inspired us further to rethink dominant XAI approaches and motivated the aim of developing an alternative approach to XAI. As a theoretical lens to help us in this endeavor, we turned to framing theory.

Hence, we are now positioned to compare the frames that guide the dominant practice of post-hoc explainability in XAI,

Table 9.2: Qualities of XAI where the target audience is domain experts

Quality	**Conventional XAI view of framing**	**Sleep experts' view of framing**
	Strategic view of framing	**Interactive view of framing**
Transparency	Transparent view of model internals.	Not necessary or plausible to make the process transparent.
Communication	Explanation as a one-way message from ML algorithm(s) to human.	Explanation as meaning-making through communication between human and ML algorithm(s), embedded in a rule-based framework.
	Object view of framing	**Process view of framing**
Model framework	Rule-based or pattern recognition.	Rule-based and pattern recognition.
Language	Simple, familiar language.	The language should use terminology from the rules.

using the literature on XAI as a proxy for this practice, with our findings of the frames that guide the domain practice we study in this chapter. We see that these practices are guided by frame differences. These frame differences are partially overlapping in practice, but distinguishable on an analytical level. We believe that these frame differences can hinder effective and creative human-AI collaboration if not handled properly. Because of these frame differences, any effort to refine the explainability of ML processes for people in human-AI collaboration is likely to only be effective for enhancing human-AI collaboration to a certain point. This is because efforts to mimic a domain-specific language of practitioners do not address the fact that ML processes are essentially statistical processes of pattern recognition in data [129].

We identify frame differences across four dimensions (Table 9.2): rule-based vs pattern recognition models, transparency of model internals, the familiarity of language, and one-way vs multi-way communication and interaction. Taken together, these frame differences motivate an alternative approach to integrating XAI in different practices.

When it comes to the model framework, the literature on XAI is built on the assumption that ML models are trained to perform pattern recognition. Indeed, a software system based on AASM rules would not require XAI techniques to explain its

decisions. However, the sleep experts expected the AI system to operate within the framework of the AASM rules, or at least give explanations based on the rules. This suggests a different view of XAI, where the AI system performs pattern recognition within a rule-based framework, similar to how the sleep experts themselves work. When it comes to interactivity and multi-way communication, there is a similar difference in framing between the XAI literature and the sleep experts. The XAI literature seems to implicitly assume that the explanation is a one-way message that the system "produces" or "provides". In contrast, the sleep experts expressed interest in asking further questions about the explanation, suggesting a more interactive view of framing.

Although the XAI literature emphasizes the importance of using familiar language in explanations, the sleep experts took this idea further, insisting that the explanation should not only use terminology directly from the AASM rules but also follow their logic. Finally, the XAI literature asserts that explanations should reflect the model's internal logic. This is in stark contrast to the opinion of many of the sleep experts, who did not emphasize the need for a "true" description of the decision-making process of the AI system. Instead, they preferred the explanation to justify the decision in terms of the AASM rules. However, some participants did express concerns about the explanation being misleading. This difference in framing suggests that the sleep experts view the explanation as a tool for making sense of the decision within the context of the AASM rules, rather than as a reflection of the AI system's internal logic.

The shift we propose in this chapter from persuasion-based XAI to co-creating meaning with AI entails two significant shifts, as outlined in Table 9.2. The first is a shift from a strategic to an interactive view on framing in human-AI collaboration. In the former, practitioners are viewed as targets of convincing explanations of ML processes, whereas in the latter, they are viewed as active participants in meaning-making together with AI decisions that signal ambiguities or even conflicts between ML algorithms and rule-based systems. The second shift is one from an object view of explanations in human-AI collaboration to a process view of framing in human-AI collaboration in which both AI agents and human practitioners play a part. In the former, explanations of ML processes are static outputs in the form of messages, whereas in the latter, they function as triggers

of human reflection. The co-creation perspective we introduce thereby addresses the limitations of current XAI methods, which often struggle to adapt explanations to specific contexts while accurately representing ML processes. The co-creation approach also situates XAI as a social problem [140], [285], stressing the significance of interactions between humans and AI in the creation of meaning.

These insights have practical implications, as they point to the fact that the design of XAI interfaces in a specific context of human-AI collaboration should avoid relying on persuasion in efforts to improve such collaboration. Rather, it should introduce design elements that signal ambiguity and trigger human thinking about what the ML response might mean. Such signals could include design elements such as, for example, highlighting areas of uncertainty [111] or disagreement between different prediction methods, such as rule-based versus pattern recognition approaches.

Two important limitations should be noted. First, the study was based on a small sample, which limits the generalizability of the findings. Second, all participating sleep technologists were already involved in the Sleep Revolution project. This may mean that the participants were more accustomed to research activities, more reflective about scoring practices, or more positively disposed toward experimental AI tools than sleep scorers in other settings.

The next chapter explores ways to design and implement AI systems that allow humans to collaborate with ML pattern recognition models within software systems based on transparent, rule-based logic. Additionally, since this study was limited to text-based explanations, techniques that rely on a visual representation of information were not included [290]. Exploring the role of visual representations in co-creating meaning with AI thus makes a ripe area for future research.

9.7 Conclusion

This chapter has introduced a framing-based understanding of XAI and its role in human-AI collaboration. With this lens, we have problematized existing assumptions of the dominant approach to post-hoc XAI and contributed an alternative approach to tackling the black box problem in human-AI collaboration. This alternative approach entails a shift from persuasion-based XAI to co-creation

with AI. We hope that this view will enrich future research into XAI as well as enhance human-AI collaboration in practice.

Chapter 10

The MARBLE framework

Rules and patterns at multiple abstraction levels

10.1 MARBLE

The preceding chapters have shown that high predictive performance is possible in AI sleep staging, but also that predictive performance alone is not enough for clinically useful AI (Chapter 4). They further showed that epoch-by-epoch staging performance does not translate directly into accurate downstream sleep metrics, and that some of those metric targets are themselves uncertain across expert human scorers (Chapters 6 and 7). Chapter 8.3 showed that sleep architecture metrics can, in some cases, be estimated directly from PSG data rather than derived through epoch-by-epoch staging. However, this introduces a new problem: although direct metric estimation may better match the clinical endpoint, it moves further away from the intermediate representations that human experts can inspect, verify, and contest. This is a problem of relevance: the most useful representation of the case depends on the level of the workflow at which the decision is being made. We have also seen that important parts of the sleep architecture analysis procedure can be translated into executable,

auditable rules (Chapter 5), while effective human-AI collaboration requires interactive systems that preserve human oversight and accountability (Chapter 9).

Together, these findings point toward a neurosymbolic form of decision support in which rule-based and pattern-based processes operate across multiple abstraction levels. We call this the Multiple Abstraction-Level Rule-Based Learning Engine (MARBLE). MARBLE is not a single algorithm or model; it is a decision support system architecture for embedding ML models within a symbolic reasoning system, in such a way that the human user's role as the final authority is preserved and the system's primary contribution is to surface where human input is most needed.

Figure 10.1 shows a schematic overview of MARBLE, applied to sleep analysis. The workflow is represented as a hierarchy of abstraction levels (dashed regions), progressing from raw PSG signals to micro-annotations, to epoch-level sleep staging, and on to derived sleep metrics and the diagnostic procedure. Each level is supported by two complementary paths: a rule-based pathway that follows the AASM scoring rules and other clinical guidelines, and a pattern-based pathway implemented as a black-box ML model trained on data. The remainder of this section expands on this framework and explains why pathways are kept in parallel at every level rather than being collapsed into a single mode of reasoning at each step.

10.1.1 Abstraction levels and pathways

We use the following working definition of an *abstraction level* in MARBLE. An abstraction level is a stage in the analysis pipeline defined by (i) a vocabulary of objects and (ii) a transformation that produces those objects from objects at the immediately lower level. Two processes operate at the same level if they consume objects from the same lower-level vocabulary and produce objects in the same level's vocabulary, regardless of whether they do so by symbolic computation or by statistical pattern recognition. Levels are arranged in a strict linear order: each level depends only on the level immediately below it. As a consequence, every output of the system at any level can in principle be traced back through a single chain of transformations to the raw input.

10.1.2 Flags as the primary output

Because MARBLE runs two pathways at multiple levels, it produces multiple values for the same quantity. At the sleep metric level, for instance, the rule-based pathway and the pattern-based pathway each produce an estimate of N3 duration. MARBLE's primary output is therefore not an estimated quantity such as a sleep stage or a metric, but a collection of warning flags raised across the abstraction hierarchy as a function of how those parallel estimates relate to one another. Two flag types are exposed to the user.

The first is an *uncertainty flag*, raised when a pattern-based model indicates low confidence in its decision. The uncertainty flag for an epoch's sleep stage is raised when the probability of the majority class falls below a threshold τ_u, currently set heuristically to 0.5. An uncertainty flag tells the user that the pattern-based pathway has weak evidence for its prediction at the level concerned, whether because of poor signal quality, ambiguous patterns, or genuinely borderline cases.

The second type of flag is a *discrepancy flag*, raised when the rule-based and pattern-based pathways at a given level produce different conclusions. For categorical outputs, such as epoch-level stage, this is straightforward: a discrepancy is a class mismatch. For real-valued quantities, such as sleep metrics, we raise the flag when the two estimates differ by more than a threshold τ_d, currently set heuristically to 10% of the rule-based value. A discrepancy flag tells the user that the two reasoning modes disagree at the level concerned and that the disagreement deserves human attention.

Both thresholds are user-tunable parameters of the system. The default values are currently chosen on heuristic grounds and represent reasonable initial values rather than empirically calibrated decision boundaries; their refinement against clinical workflow data is identified as future work in Section 11.5. In MARBLE, discrepancy and uncertainty flags are the primary mechanism by which the system indicates that human intervention is needed.

10.1.3 Why parallel pathways at every level

Running two pathways at every level may seem wasteful: why not use rules wherever they apply and fall back on pattern recognition only at the lowest level, where rules cannot reach? There are three reasons.

The first is protection from what is sometimes called the "computer says no" situation. A purely rule-based pathway can reach a confident conclusion on the basis of a single, perhaps borderline, condition deep in the rule chain, with no internal mechanism for noticing that the input was atypical. Pattern recognition at higher abstraction levels acts as a sanity check against this. An ML model trained to estimate N3 duration directly from a full-night recording, for example, may not be as accurate as summing the N3 epochs of a rule-based hypnogram, but it will produce a sensible estimate even when the rule-based pathway has gone astray on a single problematic micro-annotation. The disagreement is what alerts the system, and ultimately the user, that something has gone wrong.

The second is that uncertainty propagates through systems of rules in complex and unpredictable ways. ML models at lower levels produce probabilistic predictions, but feeding those probabilities through hard thresholds and discrete aggregation steps causes small errors to cancel out, compound, or, in some cases, drastically alter the outcome at higher levels. Maintaining a parallel pattern-based pathway at higher levels gives MARBLE a second, end-to-end estimate that is not subject to the same hard thresholds, so that the two pathways together delineate a plausible range for the quantity in question.

The third reason is that having both pathways at every level is what makes the system useful as a guide to where human review is needed. When the two pathways agree at a given level, the user can move on; when they disagree, the user knows the disagreement is local to that level and can drill into the evidence at that level rather than having to inspect the entire pipeline. In this way, the parallel structure is not just a redundancy mechanism but a tool for relevance realization in the clinical decision process [143].

10.1.4 Relation to prior hybrid and neurosymbolic systems

The wider neurosymbolic AI literature has converged in recent years on the view that pattern-based and rule-based reasoning should be *integrated* rather than merely *combined* [144]. Most existing neurosymbolic architectures, however, are designed for general-purpose reasoning tasks and do not prescribe a specific structural relationship between symbolic and neural components.

MARBLE is more constrained: it ties the integration to the abstraction hierarchy of an existing regulated workflow, and its primary output is the disagreement structure rather than a unified prediction.

The novelty of MARBLE is therefore not the mere combination of rules and learning. It is the explicit treatment of a clinical workflow as a multi-level abstraction hierarchy in which both reasoning styles operate in parallel at each level, and in which the system's primary contribution is to surface uncertainty and rule-pattern disagreement as the focus of human attention. To our knowledge, no existing system has been designed this way for clinical sleep analysis, and the broader neurosymbolic and CDSS literatures contain no architecture that combines all three criteria: multi-level hierarchy; level-wise parallelism; and disagreement-as-output, in the form proposed here.

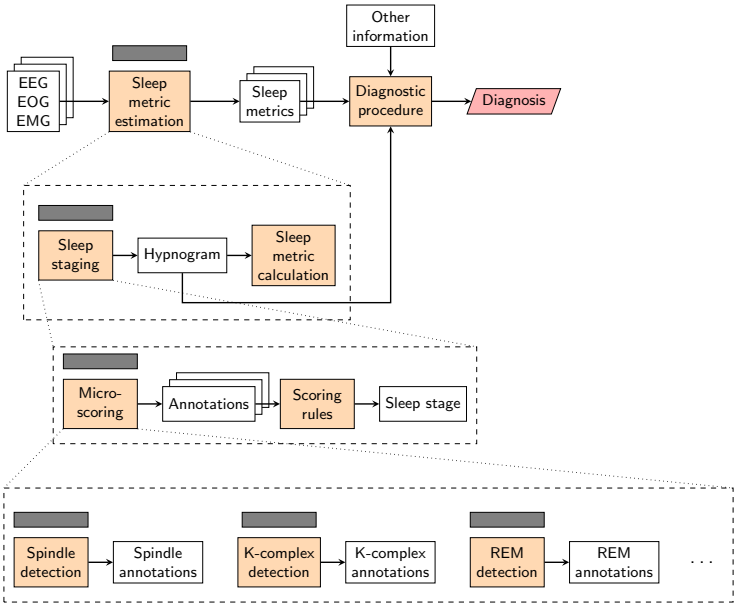


Figure 10.1: Schematic overview of the MARBLE (Multiple Abstraction-Level Rule-Based Learning Engine) framework applied to clinical sleep analysis. The workflow is represented as a hierarchy of abstraction levels (dashed regions), progressing from raw PSG signals (EEG/EOG/EMG) to low-level micro-annotations (e.g., spindles, K-complexes, REMs), to epoch-level sleep staging and the resulting hypnogram, and further to derived sleep metrics that are combined with other information in a diagnostic procedure to produce a diagnosis. At each level, MARBLE pairs a rule-based process (e.g., scoring rules and deterministic calculations) with a pattern-based black-box counterpart (black rectangles) trained on data, enabling a comparison between rules and learned statistical evidence and supporting human oversight when the two pathways disagree.

10.2 SleepScorer: MARBLE applied to sleep analysis

The MARBLE framework is a conceptual architecture for decision support systems. To see what such an architecture looks like in practice when applied to sleep analysis, we now describe a concrete implementation. SleepScorer is a decision support system for sleep architecture analysis based on MARBLE. A prototype of the system is available for use at https://sleepscorer.com.

10.2.1 System architecture

SleepScorer is implemented as a web application with a Python Django backend, the Django ORM over a relational database, and a React frontend. PSG recordings are uploaded as EDF files and loaded using MNE [204]. Signals are preprocessed and divided into 30-second epochs, and the processed data are stored in the database. The system uses the rule-based sleep staging from Paper II as its rule-based epoch pathway, and the ML models from Papers I and V as its pattern-based pathways at the epoch and sleep-metric levels, respectively. At the micro-annotation level, ML detectors for events such as sleep spindles, K-complexes, SWA, and REMs run in parallel with signal-processing rules that check the AASM-specified amplitude, frequency, and duration criteria for each event type.

10.2.2 Navigating the abstraction hierarchy

The SleepScorer user interface is organized around the abstraction hierarchy in Figure 10.1: a sleep report overview, a sleep metric view, a hypnogram view, an epoch view (where signals and micro-annotations are shown), and a micro-event view. The user navigates between these views, keeping track of their current level on the *Analysis Level* navigation bar at the top. Throughout these levels, the system presents three "paths", a pattern-based estimator (ML models), a rule-based estimator (scoring rules and other logic) and a manual override.

The user (typically a sleep technologist) begins at the sleep report level (Figure 10.2), where they see patient metadata, recording metadata, and a list of sleep metrics. With the "Show warnings" toggle enabled, sleep metrics that need attention are marked with

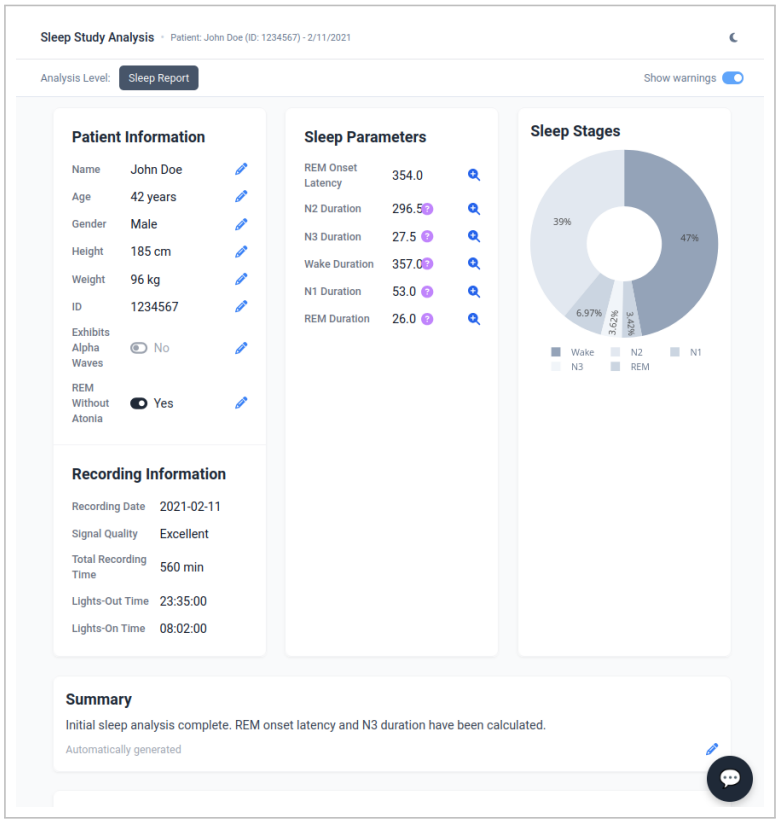


Figure 10.2: SleepScorer report-level overview showing patient and recording metadata, automatically derived sleep parameters, and stage distribution. Warning icons highlight parameters that require review due to uncertainty or disagreement between rule-based and pattern-based pathways.

warning icons, and hovering the mouse over the warnings reveals the reason for the warning.

Clicking the magnifying glass icon next to a sleep metric opens the sleep metric view (see Figure 10.3). For N3 duration, for example, the rule-based pane states the calculation explicitly: the sum of N3-epoch durations in the hypnogram, by AASM rule; while the pattern-based pane shows the estimate produced by the heuristic signal intensity (HSI) and ViT models from Paper V, together with the soft-aggregation method described there. If the result is faulty, the user can correct it directly via the manual override pane or by descending to a lower level. To do the latter, they click "View in Hypnogram".

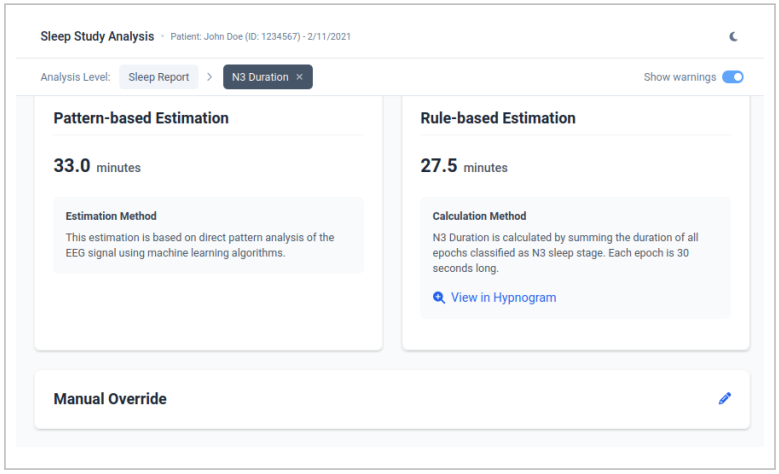


Figure 10.3: Parameter-level view for N3 duration with pattern-based and rule-based estimates shown side by side. The rule-based pane states the explicit calculation (summing N3 epochs) and links to the hypnogram. The user can manually set the N3 duration using manual override.

The hypnogram view (Figure 10.4) shows two hypnograms. One is rule-based and shows all epochs scored using the scoring rules. The other is estimated from PSG signals using the HSI or ViT models (pattern-based). Epochs are highlighted in yellow where the two pathways disagree (a discrepancy flag) and in purple where the pattern-based model is uncertain (an uncertainty flag). Clicking on a flagged epoch opens the epoch view (Figure 10.5),

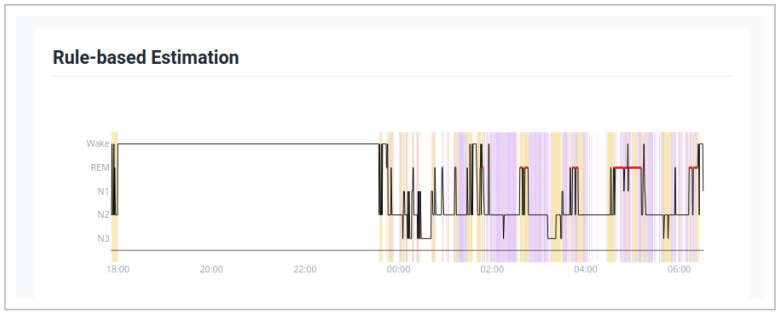


Figure 10.4: Hypnogram view in the SleepScorer prototype, designed to clarify where human oversight is most needed. The interface displays the night's hypnogram estimates (rule-based and pattern-based) and overlays warning regions on the hypnogram: **purple** marks *uncertainty* (low-confidence or weak evidence in the pattern-based estimate) and **yellow** marks *discrepancy* (conflict between rule-based and pattern-based outputs). Clicking on an epoch takes the user to the epoch-level view for that epoch.

where the PSG signals are plotted with the underlying micro-annotations. The same three panes (pattern-based, rule-based, manual override) summarize the stage decision for the epoch: the pattern-based pane reports the ML model's stage probabilities (e.g., 80% probability of N1); the rule-based pane shows the decision implied by the AASM rules together with the chain of micro-annotations from which it was derived; and the manual override pane allows the user to set the stage directly.

The user can also change the rule-based stage by editing the micro-annotations on which the rule-based decision was based, by creating, editing, or deleting an annotation. Whenever the user makes a change at any level, SleepScorer re-runs the affected parts of the analysis pipeline, propagating the change upward. A single mislabelled SWA annotation, for example, may flip an epoch's rule-based stage from N3 to REM, and if the epoch is early in the night, the change can substantially shift REM onset latency, and thereby the sleep report. In this way, the user can correct errors where they originate.

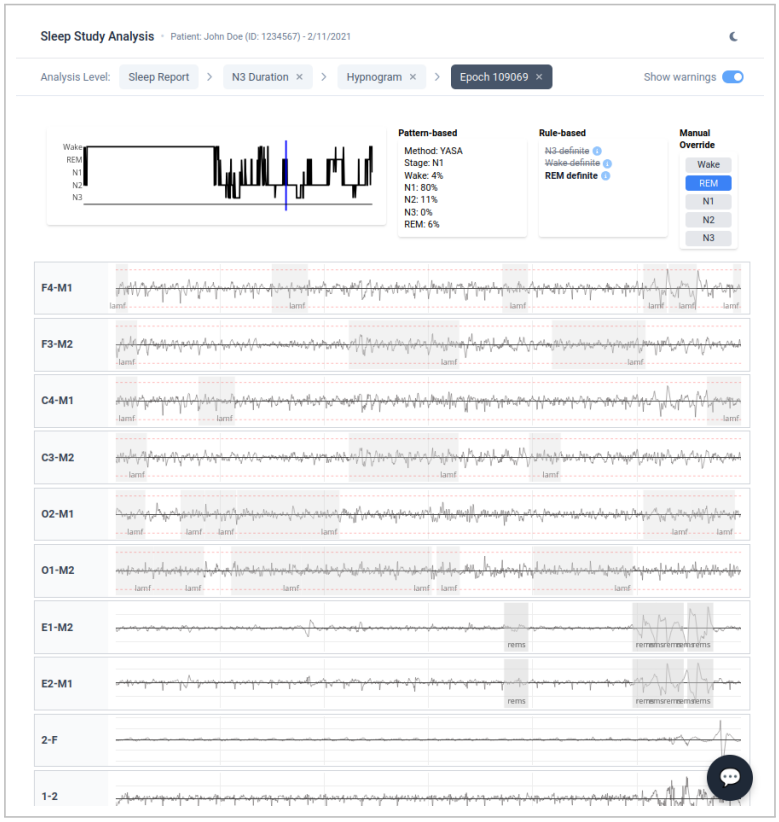


Figure 10.5: Epoch-level view in the SleepScorer prototype. The top navigation bar ("Analysis Level") shows the user's current place in the abstraction hierarchy (Sleep Report > Sleep Metric > Hypnogram > Selected Epoch). The PSG signals for the selected epoch are plotted with the micro-annotations. Three decision panes summarize the sleep stage decision for this epoch: the pattern-based pane shows the ML model predicted stage probabilities; the rule-based pane shows the stage predicted by the AASM rules; and the manual override pane allows the user to set the epoch stage directly. This example illustrates a rule/pattern discrepancy, pattern-based suggests N1 while the rules say REM, resolved via a manual override. Higher-level quantities are then recomputed and propagated upwards.

10.2.3 The chat assistant

At every analysis level, SleepScorer provides an embedded chat interface (Figure 10.6) in which the user can interact with an assistant in natural language. The assistant does not generate post-hoc explanations independently of the rule-based pathway. The natural-language explanation mechanism developed in Paper II produces a structured trace of the rule-based decision in clinical language, and this trace is the only substantive input the assistant receives about the recording in the current view. The assistant's role is to render that trace conversationally, to map user questions to the relevant entries in the trace, and to clarify the meaning of clinical terms. At the epoch level, for instance, a user asking why an epoch was not classified as N2 is answered by reference to the specific necessary criteria in the trace for which evidence was lacking, such as an absent K-complex or spindle.

This design constrains the assistant to operate within MARBLE's commitments rather than against them. It does not eliminate the failure modes of LLM-based systems entirely: the assistant can still misrender or omit parts of a long trace, and its phrasings should be treated as an aid to navigation rather than as a primary source of clinical justification. The text rendered in the rule-based pane of the relevant view remains the authoritative explanation.

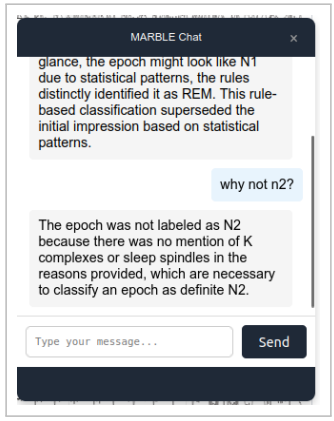


Figure 10.6: Chat interface for querying decisions in the SleepScorer system. Answers are grounded in the scoring rules and available micro-annotations.

10.2.4 Known limitations

Several limitations of the current SleepScorer prototype are worth noting. The first concerns alert fatigue, a well-documented failure mode of clinical decision support systems in which an excess of warnings causes users to ignore them or override them by reflex [314]. A closely related issue is the potential overreliance on alerts [315]. The number of flags raised during a typical recording depends on the thresholds τ_u and τ_d, which are currently set on heuristic grounds. Calibrating these thresholds against clinical workflow data, so that the rate of raised flags is informative rather than overwhelming, is necessary before SleepScorer is deployed beyond its current research setting.

The second concerns the agreement-on-error case, in which the rule-based and pattern-based pathways agree but are both wrong. MARBLE raises no flag in this case, and its design implicitly assumes that the two pathways fail independently. This assumption is plausible because the two pathways are constructed from different sources of information — one from explicit rules over micro-annotations, the other from learned statistical patterns over raw signals — but it is an assumption nonetheless. Quantifying the rate of joint failure on held-out data is identified as future work in Section 11.5.

The third concerns calibration of borderline cases at the micro-annotation level. When the underlying micro-annotations are themselves uncertain, the rule-based pathway's downstream conclusion inherits that uncertainty in ways the current implementation does not fully propagate. A more thorough propagation of uncertainty from the micro-annotation level upward through the rule-based pathway is a natural next step.

The fourth, already noted above, concerns the chat assistant. Even with grounding in the rule trace from Paper II, the assistant remains an LLM and inherits the failure modes of LLMs; users should treat its output as a conversational view onto the underlying trace rather than as an independent justification.

Fifth, the SleepScorer system has not been evaluated empirically. The prototype is presented here as a design artifact intended to demonstrate that the MARBLE framework can be instantiated end-to-end, with formal usability and clinical validation studies left to future work.

Overall, SleepScorer illustrates what a rule-governed AI decision

support system can look like in practice. It is not based on a single perfect model, but a decision support system that (1) maintains a clear separation between statistical pattern recognition and rule-based logic, (2) surfaces discrepancies between the two modes, and (3) allows the user to make corrections at the appropriate level.

Chapter 11

Discussion

The results of this thesis span six papers that address distinct technical questions: architecture search, rule coding, simulation of measurement error, regression modeling, and XAI. Taken together, however, they converge on the overarching thesis questions: in PSG sleep architecture analysis, what decisions should be automated and how should they be justified? The contributions collectively reveal a set of tensions: between statistical performance of ML models and clinical reasoning; between ML evaluation metrics and clinical relevance; and between XAI explanations and clinical justifications.

11.1 Key findings

In Paper I, we used neural architecture search to demonstrate how far pattern-based learning can be pushed for automatic sleep staging. Given sufficient data and computational power, we can find and train ML model architectures that learn to recognize statistical patterns in PSG signals, resulting in a highly accurate sleep staging model. However, this is achieved without any reference to the clinical framework of rules and guidelines. The model is optimized solely for predictive performance. The predictions made by the model are not guaranteed (and in fact highly unlikely) to operate on the basis of AASM scoring rules.

In Paper II, we approached the same problem from the opposite direction. Instead of training a model to recognize patterns, we hard-coded the AASM scoring rules into executable logic. This

produced a fully transparent staging model, where every decision is justified by faithful reference to the scoring rules, in terminology familiar to clinical staff. But the transparency comes at the cost of performance. The algorithm's accuracy is lower than DL models such as the one we trained in Paper I. The shortcomings are not arbitrary; the algorithm struggles to classify scenarios where the scoring rules are ambiguous, such as for epochs of N1 or wake. Sleep staging requires more than just following the scoring manual. Human scorers use their judgment, common sense, and pattern recognition to navigate ambiguity in the rules.

Paper III reframes the task of automatic sleep analysis by calling into question the dominance of sleep staging ML. The decision to focus ML research in sleep architecture analysis on sleep staging classification has been shaped by historical and practical reasons, such as the availability of training data with epoch-level labels and the desire to emulate a time-consuming step in the manual workflow. But our simulation results show that this choice of abstraction level has consequences because the mapping of staging errors to derived sleep metrics is non-linear, complex, and often counterintuitive. The paper shows that prevailing evaluation metrics of sleep staging ML, accuracy, and Cohen's kappa, are weak indicators of performance in sleep metric estimation and hold limited clinical relevance. The paper motivates a shift in abstraction levels, presenting sleep metric estimation as an obvious candidate.

Paper IV strengthens this point by showing that human agreement at the level of epoch-by-epoch staging does not translate uniformly into agreement at the level of sleep metrics. Some sleep metrics, such as TST, SOL, WASO, and SE, were reliable across scorers, whereas others (especially metrics describing how sleep is divided across NREM stages) were much less reliable. This leaves human consensus sleep metrics as the most defensible proxy for a "ground-truth" value.

Paper V builds on this finding by asking whether sleep metrics should be estimated more directly from PSG data rather than derived through a hypnogram. The paper investigates methods that reduce or bypass reliance on the hypnogram, including soft aggregation of hypnodensity and direct sleep metric estimation. Because Paper IV showed that scorer-derived metric values vary meaningfully, Paper V evaluates these methods against consensus-based sleep metric references rather than treating a single scorer-

derived value as ground truth. Under that framing, some of the proposed methods achieve errors comparable to, and for some metrics lower than, those of individual human scorers in leave-one-out comparison against consensus.

This is encouraging, but also conceptually disturbing: should we start using sleep metric estimation models as the "ground-truth" if they are shown to consistently outperform human scorers? Clinical sleep analysis can not be reduced to producing a handful of summary metrics. Sleep metrics are a part of a broader system of interpretation and decision-making, where human common sense and intuition remain essential.

Paper VI reinforces this point by showing qualitatively that training a high-accuracy black-box model that outputs convincing explanations fails to solve the underlying issue. Interviews with sleep technologists indicate that they expect AI to operate within a framework of AASM rules and express decisions in familiar terminology. They want to be able to examine decisions made by AI, interactively querying and probing explanations made by the system.

Together, these results sharpen the chapter-opening thesis questions of what to automate and how to justify the result into more concrete design decisions: where in the hierarchy of clinical sleep analysis AI should be put, and when at that level it should be symbolic and when it should be neural.

The dominant choice in the literature is to apply AI at the epoch-by-epoch staging level. The work in this thesis suggests that this choice is problematic and that moving upward in the hierarchy, to sleep metric estimation, is more closely aligned with clinical practice. But the hierarchy continues: why not automate the creation of sleep reports directly from PSG signals? Should diagnosis be automated end-to-end? At whatever level is chosen, neither symbolic logic alone nor neural pattern recognition alone is sufficient. A rule-based pathway loses performance where the scoring manual is ambiguous (Paper II), while a pure pattern-based system loses alignment with clinical rules (Papers I and V).

These design decisions are coupled. The abstraction level imposes limitations on the available representations and on the ways they can be evaluated (Paper III), and the choice of reasoning mode interacts with the level. Rather than selecting a single level and a single mode of reasoning, we propose a framework that runs symbolic and neural processes in parallel at multiple levels of

abstraction and uses their agreement (or disagreement) to guide human-AI collaboration.

11.2 Answering the research questions

Returning to the four research questions stated in Section 1.1, the contributions of the thesis support the following answers.

> **RQ 1** *Under what conditions do conventional ML sleep staging models and evaluation metrics fail to be clinically useful?* The thesis shows empirically that conventional ML sleep staging models and the evaluation metrics (such as epoch-level accuracy and Cohen's kappa) are insufficient in any practical sleep architecture analysis. A good epoch-by-epoch staging agreement is no guarantee of adequate predictive strength in sleep architecture metrics.
>
> **RQ 2** *Are there targets and evaluation metrics that are more robust and clinically relevant than epoch-by-epoch sleep stage classification if the clinical task is interpreting sleep architecture?* The thesis shows that more clinically relevant targets exist, especially approaches that move beyond the hypnogram as an intermediate representation and instead estimate sleep architecture metrics directly.
>
> **RQ 3** *To what extent can clinical sleep scoring rules be translated into an executable procedure, and where do ML models provide support within that procedure?* The thesis shows that large parts of the clinical sleep scoring rules can be translated into executable code. At the same time, it also shows that these must be run in parallel with ML because real signals are noisy and the rules are ambiguous.
>
> **RQ 4** *What must a clinical sleep architecture analysis decision support system provide for human experts to collaborate with AI models?* The thesis shows that a clinically useful AI decision support system must do more than return predictions and static explanations. It must support interaction, convey uncertainty, and explain how evidence from the input data is used by the rules of the procedure. In particular, it must make visible where rule-based and pattern-based reasoning are in agreement and where they diverge. In this

way, the human user remains the accountable decision maker, while the system is used to navigate the multi-level decision process.

Thesis questions *In PSG sleep architecture analysis, what decisions should be automated with ML, and how should those decisions be justified?* The thesis answers these overarching questions by proposing that ML automation in PSG sleep architecture analysis should occur at multiple abstraction levels in the analysis hierarchy. The justification of decisions should be grounded in the rules governing the workflow, with ML contributing pattern recognition inside the wider decision support architecture, rather than providing explanations or justifications.

11.3 Limitations

The work presented in this thesis has some significant limitations. In addition to the limitations listed in individual chapters (Chapters 4-9), the thesis results are limited in at least three areas.

Firstly, all empirical results in the thesis rely on three datasets (SHHS, SleepEDFx, and SRMS) and on the populations and recording conditions of those datasets. SHHS is large but draws from an older, predominantly community-dwelling cohort and uses recording protocols that differ from contemporary practice. SleepEDFx is small, and is scored under the R&K manual rather than the AASM rules. SRMS comprises only 50 full-night PSG recordings. Findings about inter-scorer reliability (Paper IV), the rule-based system's behavior on real signals (Paper II), and direct sleep metric estimation under scorer disagreement (Paper V) should therefore be read as descriptive of these conditions rather than as universal claims. The generalizability of the results and proposed methods is thus potentially limited, as shown in Paper V when the performance of models trained on SHHS was severely degraded when they were applied to SRMS. The rule-based pathway of Paper II has been evaluated only on SRMS and the MARBLE architecture of Chapter 10 inherits these limitations.

Secondly, the MARBLE and the SleepScorer prototype have not undergone any clinical evaluation. SleepScorer is presented in Chapter 10 as a design artifact intended to demonstrate that the MARBLE framework can be implemented end-to-end. Whether

the parallel rule-based and pattern-based pathways actually support clinical decision making, whether the flags they raise are informative without being overwhelming, whether the explanations are accepted by sleep technologists as relevant, whether unflagged predictions are reliable enough to act on, and whether the system saves time without degrading quality are all open questions. The directions outlined in Section 11.5 are aimed at addressing them.

Thirdly, several of the empirical findings in the thesis are evaluated against expert consensus. Paper IV shows that consensus is not a fully stable reference at the level of individual sleep metrics, but the rest of the thesis continues to rely on it because no better reference is available. Deeper validation of any of the proposed methods, including SleepScorer, will ultimately require evaluation against external clinical outcomes rather than against scorer opinion. There is also a circularity concern that is worth naming: the same expert population that produces the consensus also produces the labels on which the pattern-based models are trained, so claims that a model approaches or exceeds consensus performance are claims about agreement with a particular labeling tradition rather than about physiological accuracy.

11.4 Looking beyond sleep

Sleep is far from the only domain where the integration challenges described above arise, and the MARBLE framework is intended to be useful beyond sleep medicine.

For the MARBLE framework to be applicable, three properties must hold. (1) The task must be a multi-level decision procedure (see Figure 11.1) in which high-level decisions are composed of lower-level sub-decisions. (2) At each level, there must be sufficiently explicit rules to support a rule-based pathway, with intermediate artifacts that are inspectable and auditable. And (3) there must be enough data and structure at each level to make a complementary pattern-based pathway worthwhile, so that the two pathways can be run in parallel and their disagreement used as a signal for human oversight. Sleep architecture analysis fulfills all three criteria. The AASM scoring manual codifies rules at all levels; the workflow has explicit intermediate artifacts at multiple abstraction levels, from micro-events through hypnograms and sleep metrics to the clinical report; and large labeled PSG datasets

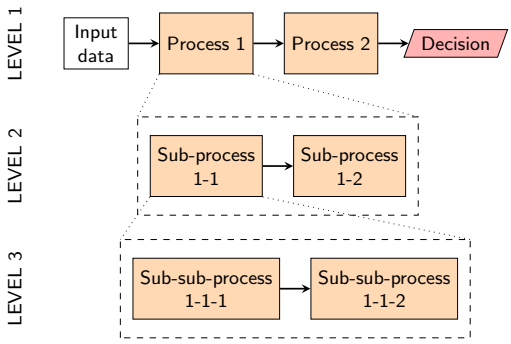


Figure 11.1: A rule-based decision-making procedure, consisting of processes at multiple abstraction levels. Process 1 consists of sub-processes 1-1 and 1-2, which in turn consist of sub-sub-processes.

exist to train pattern-based models.

Many regulated domains share these properties to varying degrees. In medicine, radiology reporting has codified guidelines and a hierarchy of intermediate artifacts. In education, the assessment of students has rubrics and guidelines that function as rules. In finance, credit underwriting and mortgage decisions sit on top of regulations and procedures and are increasingly served by ML-based credit scoring.

To illustrate what an implementation of MARBLE in another domain could look like, Figure 11.2 sketches a hypothetical MARBLE instance for mortgage application review. Mortgage decisions are suitable as an illustrative case because they are regulated and depend on both predictive assessments and explicit decision rules. A credit score may summarize borrower risk, but it is only one component in a wider decision procedure that also includes payment ability, collateral, loan-to-value (LTV) constraints, documentation checks, Central Bank rules, and institution-specific policy rules [316]–[320].

In a hypothetical MARBLE mortgage system, rule-based procedures would produce the auditable decision path, while ML models would run in parallel to flag discrepancies, uncertainty, or cases requiring human review. The system would therefore use ML to support the regulated procedure, rather than replace it.

Figures 11.3 and 11.4 show how this could appear in a decision

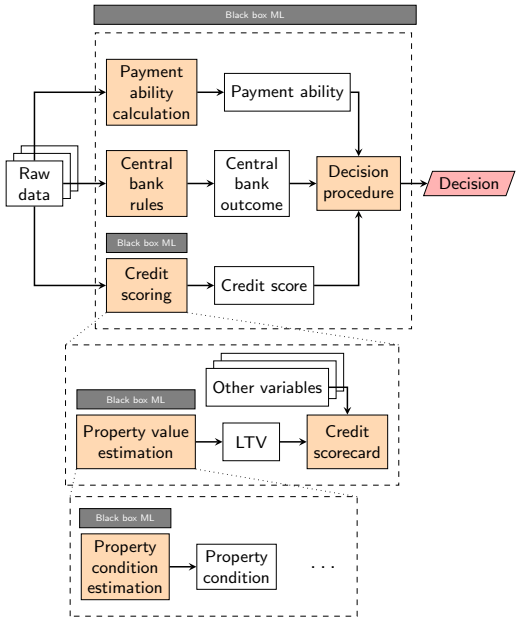


Figure 11.2: Multi-level mortgage decision pipeline for a financial institution in Iceland. Raw application data and documentation are transformed into intermediate variables and sub-assessments, which feed an explicit rule-based decision procedure. Pattern-based ML models provide parallel estimates at the same abstraction levels.

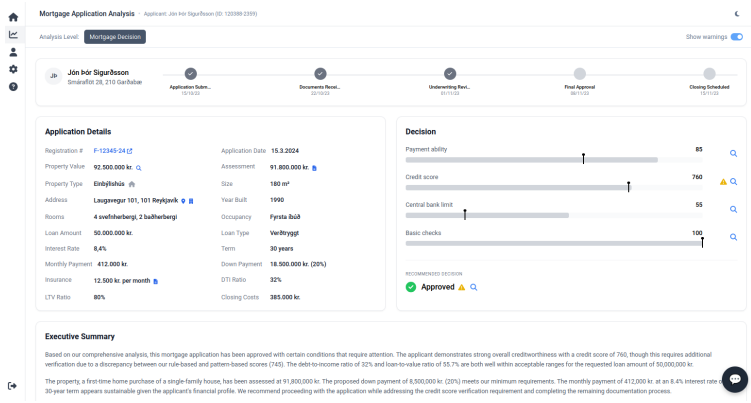


Figure 11.3: Mortgage application overview in the demonstration interface, showing the application summary, current recommendation, main decision components, and warnings requiring review.

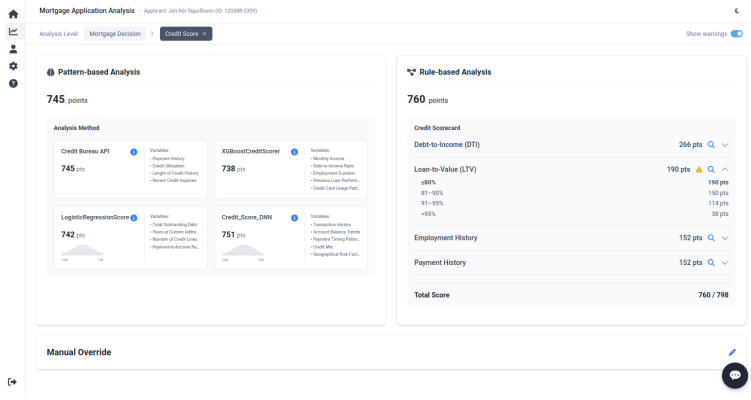


Figure 11.4: Credit score view in the prototype interface. The pattern-based analysis shows ML credit-score estimates, while the rule-based analysis shows the scorecard and its supporting evidence. Warning indicators flag discrepancies at lower abstraction levels, such as LTV.

support interface.[1] The analyst begins with an application-level overview and can drill down into any flagged component. At each level, the rule-based pathway shows the formal basis for the current value or decision, while the pattern-based pathway provides an independent estimate. Disagreements between the two are surfaced as warnings, and the user can investigate or override the system.

This mortgage example is only a demonstration. It does not implement a validated scorecard or evaluate real mortgage decisions. Its purpose is more limited: to show that the MARBLE framework (multiple abstraction levels, parallel rule-based and pattern-based pathways, drill-down on discrepancies, and human override) can be expressed in a regulated domain that is structurally different from sleep medicine.

11.5 Future work

This thesis opens several directions for future work. The SleepScorer prototype must be evaluated under conditions that resemble clinical practice. A study with two complementary parts would address most of the open questions identified in Section 11.3. The first part could be a retrospective evaluation of SleepScorer outputs against multi-scorer reference data already collected within the Sleep Revolution project. This would make it possible to evaluate the system's predictive accuracy both at the level of sleep stages and, more importantly, at the level of sleep architecture metrics, against consensus-based reference values rather than against a single scorer's labels. The retrospective evaluation would also support targeted analyses that the prototype's design assumptions depend on but that have not yet been validated: that the epochs flagged by MARBLE's parallel pathways correspond predomi to epochs where human scorers also disagree; that unflagged epochs are reliably correct (the agreement-on-error case discussed in Section 10.2.4); and that the flag rate produced by the current heuristic thresholds is informative rather than overwhelming once calibrated against scorer-disagreement data. The second part would be a small prospective workflow study in which experienced sleep technologists score a fixed set of recordings under two conditions, with and without SleepScorer assistance. The relevant outcome measures are time-to-score, agreement of the final scoring

[1] There is a recording of the UI at https://youtu.be/6edgoRQUmbw.

with the multi-scorer consensus, willingness to act on flags, the proportion of system-suggested revisions accepted by the user, and qualitative feedback on the explanations and on the chat assistant.

Together, these two parts would test whether MARBLE's design assumptions hold in practice and whether the system, as currently implemented, produces a workflow that experts find usable and trustworthy. Particular attention should be paid to the unflagged cases in both parts of the study, since these are presented to the user as safe to trust without further scrutiny and are therefore the cases where any silent failure of the system would do the most damage.

The broader claim of the thesis, that the MARBLE framework is transferable to regulated decision support more generally, has not been tested. Implementing MARBLE in a second domain, with parallel rule-based and pattern-based pathways operating at multiple abstraction levels and the same flag-driven model of human oversight, would either strengthen the cross-domain claim or expose dependencies on the specifics of sleep medicine. Suitable candidate domains include clinical decision support outside of sleep. The mortgage application demonstration described in Section 11.4 sketches what such an instance could look like outside the clinical setting.

Future work should also explore clinical outcome validation of alternative sleep metrics. Papers III and IV show that some conventional sleep architecture metrics have substantial inter-scorer variability and may not reliably reflect the physiological quantities they are taken to measure. This raises a question that the present thesis does not answer: whether alternative sleep descriptors, for example, the heuristic signal intensity (HSI) metrics introduced in Paper V, predict clinically meaningful outcomes better than conventional hypnogram-derived metrics. A useful direction is to replicate published studies that have used conventional metrics in outcome prediction, for example, the work of Javaheri *et al.* [251] on the association between N3 sleep and hypertension, on open datasets such as SHHS, and to test whether HSI-based or direct-estimation metrics produce stronger or more reliable associations. Clinical outcomes provide an external source of truth that scorer consensus does not.

Three further directions are worth listing briefly. The path-dependent sleep metrics not covered by Paper V, sleep onset latency and REM onset latency in particular, deserve their own treatment,

since they may behave differently under direct estimation or uncertainty propagation than the additive duration metrics studied so far. And the micro-annotation models used in Paper II should be replaced by separately validated event-level detectors, against expert event-level annotations, before the rule-based pathway is treated as a primary sleep stage scoring method rather than as an auditable cross-check.

Chapter 12

Conclusion

This thesis investigated how AI and ML can be integrated into the regulated workflow of polysomnography-based sleep architecture analysis without sacrificing the justifiability of outcomes. The argument developed across the six papers and synthesized in the MARBLE framework of Chapter 10 is that AI integration in clinical sleep analysis should not be framed as the automation of a single prediction task, typically epoch-by-epoch staging, but as the careful design of decision support across a rule-governed hierarchy. Workflows should be implemented as multi-level executable emulations of the rules, ML models should be inserted in parallel at each level rather than concentrated at one, and discrepancies between the rule-based and pattern-based pathways should be used to guide human oversight. The human expert remains the accountable decision maker.

Beyond this integration argument, the thesis produced several specific contributions. Paper I introduced a data-local autonomous LLM-guided neural architecture search framework for multimodal multiclass classification, demonstrating that competitive sleep staging models can be discovered automatically under data-locality constraints typical of clinical settings. Paper II contributed, to our knowledge, the first automatic sleep staging algorithm based directly on the AASM scoring manual, showing that the task can be automated in a way that remains faithful to the scoring rules rather than learning them indirectly through human labels. Papers III and IV established, first by simulation and then empirically, that strong epoch-level staging agreement does not guarantee reliable

sleep metric estimation, challenging an assumption that has been implicit in much of the automatic sleep staging literature. Paper V developed direct sleep metric estimation methods that bypass the hypnogram and showed, on N3 duration, that a low-capacity model can exceed the performance of some individual human scorers when evaluated against expert consensus. Paper VI established design requirements for human–AI collaboration in this setting, and SleepScorer in Chapter 10 instantiated those requirements as a working prototype.

Although the work is rooted in sleep medicine, the problem it addresses, i.e., how to integrate AI into a regulated workflow without losing the chain of justification that makes the workflow defensible, is shared by many other domains. The MARBLE framework is offered as a candidate design principle for that broader problem. Whether it generalizes beyond sleep is a question for the work that follows from this thesis rather than one settled by it.

Bibliography

[1] O. Ismail and N. Ahmad, "Ethical and Governance Frameworks for Artificial Intelligence: A Systematic Literature Review", en, *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 19, no. 14, pp. 121–136, Jul. 2025, ISSN: 1865-7923. DOI: 10.3991/ijim.v19i14.56981. [Online]. Available: https://online-journals.org/index.php/i-jim/article/view/56981.

[2] T. Nguyen and A. Elbanna, "Understanding Human-AI Augmentation in the Workplace: A Review and a Future Research Agenda", en, *Information Systems Frontiers*, Mar. 2025, ISSN: 1572-9419. DOI: 10.1007/s10796-025-10591-5. [Online]. Available: https://doi.org/10.1007/s10796-025-10591-5.

[3] W. L. Currie, J. M. Leimeister, D. Schlagwein, and L. Willcocks, "Rethinking technology regulation in the age of AI risks", en, *Journal of Information Technology*, vol. 40, no. 3, pp. 236–245, Sep. 2025, ISSN: 0268-3962, 1466-4437. DOI: 10.1177/02683962251378815. [Online]. Available: https://journals.sagepub.com/doi/10.1177/02683962251378815.

[4] A. V. Benjafield, N. T. Ayas, P. R. Eastwood, *et al.*, "Estimation of the global prevalence and burden of obstructive sleep apnoea: A literature-based analysis", eng, *The Lancet. Respiratory Medicine*, vol. 7, no. 8, pp. 687–698, Aug. 2019, ISSN: 2213-2619. DOI: 10.1016/S2213-2600(19)30198-5.

[5] C. M. Morin and D. C. Jarrin, "Epidemiology of Insomnia: Prevalence, Course, Risk Factors, and Public Health Burden", eng, *Sleep Medicine Clinics*, vol. 17, no. 2, pp. 173–191, Jun. 2022, ISSN: 1556-4088. DOI: 10.1016/j.jsmc.2022.03.003.

[6] A. Broström, Z. Alimoradi, J. Lind, M. Ulander, F. Lundin, and A. Pakpour, "Worldwide estimation of restless legs syndrome: A systematic review and meta-analysis of prevalence in the general

adult population", eng, *Journal of Sleep Research*, vol. 32, no. 3, e13783, Jun. 2023, ISSN: 1365-2869. DOI: 10.1111/jsr.13783.

[7] R. B. Berry, *Fundamentals of sleep medicine*, eng. Philadelphia, PA: Elsevier/Saunders, 2012, ISBN: 978-1-4377-0326-9.

[8] M. M. Troester, S. F. Quan, A. A. o. S. Medicine, and R. B. Berry, *The AASM Manual for the Scoring of Sleep and Associated Events, Version 3*, en. American Academy Of Sleep Medicine, Jun. 2023, ISBN: 978-0-9706137-1-4.

[9] H. Phan and K. Mikkelsen, "Automatic sleep staging of EEG signals: Recent development, challenges, and future directions", en, *Physiological Measurement*, vol. 43, no. 4, 04TR01, Apr. 2022, ISSN: 0967-3334. DOI: 10.1088/1361-6579/ac6049. [Online]. Available: https://dx.doi.org/10.1088/1361-6579/ac6049.

[10] J. Mallett and E. S. Arnardottir, "Improving Machine Learning Technology in the Field of Sleep", en, *Sleep Medicine Clinics*, Measuring Sleep, vol. 16, no. 4, pp. 557–566, Dec. 2021, ISSN: 1556-407X. DOI: 10.1016/j.jsmc.2021.08.003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1556407X2100059X.

[11] J. Fischer, Z. Dogas, C. L. Bassetti, *et al.*, "Standard procedures for adults in accredited sleep medicine centres in Europe", eng, *Journal of Sleep Research*, vol. 21, no. 4, pp. 357–368, Aug. 2012, ISSN: 1365-2869. DOI: 10.1111/j.1365-2869.2011.00987.x.

[12] A. Malhotra, M. Younes, S. T. Kuna, *et al.*, "Performance of an automated polysomnography scoring system versus computer-assisted manual scoring", eng, *Sleep*, vol. 36, no. 4, pp. 573–582, Apr. 2013, ISSN: 1550-9109. DOI: 10.5665/sleep.2548.

[13] C. Iber, S. Ancoli-Israel, A. Chesson, and S. Quan, *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology, and Techinical Specifications, 1st ed.* 2007.

[14] A. Rechtschaffen and A. Kales, *A manual of standardized terminology, techniques and scoring system of sleep stages in human subjects*. Los Angeles: University of California, Brain Information Service/Brain Research Institute, 1968.

[15] N. Stanley, "The Future of Sleep Staging, Revisited", English, *Nature and Science of Sleep*, vol. 15, pp. 313–322, May 2023. DOI: 10.2147/NSS.S405663.

[16] R. Ferri, P. Ferri, R. M. Colognola, M. A. Petrella, S. A. Musumeci, and P. Bergonzi, "Comparison between the results of an automatic and a visual scoring of sleep EEG recordings", eng, *Sleep*, vol. 12, no. 4, pp. 354–362, Aug. 1989, ISSN: 0161-8105.

[17] S. Nikkonen, P. Somaskandhan, H. Korkalainen, *et al.*, "Multicentre sleep-stage scoring agreement in the Sleep Revolution project", *Journal of Sleep Research*, vol. 33, no. 1, e13956, Feb. 2024, ISSN: 0962-1105. DOI: 10.1111/jsr.13956. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10909532/.

[18] S. Sigurdardottir, H. Pitkänen, H. Korkalainen, *et al.*, "Validation of manually scored multichannel frontal electroencephalography against polysomnography in a paediatric cohort", en, *Journal of Sleep Research*, vol. 34, no. 6, e70012, 2025, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jsr.70012, ISSN: 1365-2869. DOI: 10.1111/jsr.70012. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/jsr.70012.

[19] H. Alsolai, S. Qureshi, S. M. Z. Iqbal, *et al.*, "A Systematic Review of Literature on Automated Sleep Scoring", *IEEE Access*, vol. 10, pp. 79 419–79 443, 2022, Conference Name: IEEE Access, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2022.3194145. [Online]. Available: https://ieeexplore.ieee.org/document/9841539.

[20] J. Hasan, "Past and Future of Computer-Assisted Sleep Analysis and Drowsiness Assessment:" en, *Journal of Clinical Neurophysiology*, vol. 13, no. 4, pp. 295–313, Jul. 1996, ISSN: 0736-0258. DOI: 10.1097/00004691-199607000-00004. [Online]. Available: http://journals.lww.com/00004691-199607000-00004.

[21] T. Penzel and R. Conradt, "Computer based sleep recording and analysis", en, *Sleep Medicine Reviews*, vol. 4, no. 2, pp. 131–148, Apr. 2000, ISSN: 10870792. DOI: 10.1053/smrv.1999.0087. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1087079299900874.

[22] P. Anderer, G. Gruber, S. Parapatics, *et al.*, "An E-Health Solution for Automatic Sleep Classification according to Rechtschaffen and Kales: Validation Study of the Somnolyzer 24 × 7 Utilizing the Siesta Database", *Neuropsychobiology*, vol. 51, no. 3, pp. 115–133, May 2005, ISSN: 0302-282X. DOI: 10.1159/000085205. [Online]. Available: https://doi.org/10.1159/000085205.

[23] P. Anderer, A. Moreau, M. Woertz, *et al.*, "Computer-assisted sleep classification according to the standard of the American Academy of Sleep Medicine: Validation study of the AASM version of the Somnolyzer 24 × 7", eng, *Neuropsychobiology*, vol. 62, no. 4, pp. 250–264, 2010, ISSN: 1423-0224. DOI: 10.1159/000320864.

[24] I. Boira, V. Esteban, J. N. Sancho-Chust, *et al.*, "Validation of the Somnolyzer 24×7 automatic scoring system in children with suspected obstructive sleep apnea", English, *Frontiers in Medicine*, vol. 12, Jun. 2025, ISSN: 2296-858X. DOI: 10.3389/fmed.2025.1617530. [Online]. Available: https://www.frontiersin.org/journals/medicine/articles/10.3389/fmed.2025.1617530/full.

[25] S.-F. Liang, C.-E. Kuo, Y.-H. Hu, and Y.-S. Cheng, "A rule-based automatic sleep staging method", *Journal of Neuroscience Methods*, vol. 205, no. 1, pp. 169–176, Mar. 2012, ISSN: 0165-0270. DOI: 10.1016/j.jneumeth.2011.12.022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016502701100759X.

[26] L. Fiorillo, A. Puiatti, M. Papandrea, *et al.*, "Automated sleep scoring: A review of the latest approaches", *Sleep Medicine Reviews*, vol. 48, p. 101 204, Dec. 2019, ISSN: 1087-0792. DOI: 10.1016/j.smrv.2019.07.007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1087079218301746.

[27] O. Faust, H. Razaghi, R. Barika, E. J. Ciaccio, and U. R. Acharya, "A review of automated sleep stage scoring based on physiological signals for the new millennia", *Computer Methods and Programs in Biomedicine*, vol. 176, pp. 81–91, Jul. 2019, ISSN: 0169-2607. DOI: 10.1016/j.cmpb.2019.04.032. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169260718313865.

[28] R. Boostani, F. Karimzadeh, and M. Nami, "A comparative review on sleep stage classification methods in patients and healthy individuals", *Computer Methods and Programs in Biomedicine*, vol. 140, pp. 77–91, Mar. 2017, ISSN: 0169-2607. DOI: 10.1016/j.cmpb.2016.12.004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169260716308276.

[29] A. Supratak, H. Dong, C. Wu, and Y. Guo, "DeepSleepNet: A Model for Automatic Sleep Stage Scoring based on Raw Single-Channel EEG", 2017, Version Number: 2. DOI: 10.48550/ARXIV.1703.04046. [Online]. Available: https://arxiv.org/abs/1703.04046.

[30] 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, Mar. 2019, arXiv:1809.10932 [cs,

eess, stat], ISSN: 1534-4320, 1558-0210. DOI: 10.1109/TNSRE.2019.2896659. [Online]. Available: http://arxiv.org/abs/1809.10932.

[31] L. Fiorillo, G. Monachino, J. van der Meer, *et al.*, "U-Sleep's resilience to AASM guidelines", en, *npj Digital Medicine*, vol. 6, no. 1, pp. 1–9, Mar. 2023, Number: 1, ISSN: 2398-6352. DOI: 10.1038/s41746-023-00784-0. [Online]. Available: https://www.nature.com/articles/s41746-023-00784-0.

[32] H. Park, K. Park, and D.-U. Jeong, "Hybrid neural-network and rule-based expert system for automatic sleep stage scoring", in *Proceedings of the 22nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (Cat. No.00CH37143)*, vol. 2, Jul. 2000, 1316–1319 vol.2. DOI: 10.1109/IEMBS.2000.897979. [Online]. Available: https://ieeexplore.ieee.org/document/897979.

[33] J. Hwang, T. Lee, H. Lee, and S. Byun, "A Clinical Decision Support System for Sleep Staging Tasks With Explanations From Artificial Intelligence: User-Centered Design and Evaluation Study", EN, *Journal of Medical Internet Research*, vol. 24, no. 1, e28659, Jan. 2022, Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research. DOI: 10.2196/28659. [Online]. Available: https://www.jmir.org/2022/1/e28659.

[34] L. Larsen and D. Walter, "On automatic methods of sleep staging by EEG spectra", en, *Electroencephalography and Clinical Neurophysiology*, vol. 28, no. 5, pp. 459–467, May 1970, ISSN: 00134694. DOI: 10.1016/0013-4694(70)90271-3. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/0013469470902713.

[35] R. Baumgart-Schmitt, W. Herrmann, and R. Eilers, "On the Use of Neural Network Techniques to Analyze Sleep EEG Data", en, *Neuropsychobiology*, vol. 37, no. 1, pp. 49–58, 1998, ISSN: 0302-282X, 1423-0224. DOI: 10.1159/000026476. [Online]. Available: https://www.karger.com/Article/FullText/26476.

[36] Y. Hashizume, H. Kuwahara, N. Uchimura, *et al.*, "Examination of accuracy of sleep stages by means of an automatic sleep analysis system 'Sleep Ukiha'", en, *Psychiatry and Clinical Neurosciences*, vol. 55, no. 3, pp. 199–200, Jun. 2001, ISSN: 1323-1316, 1440-1819. DOI: 10.1046/j.1440-1819.2001.00824.x. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1046/j.1440-1819.2001.00824.x.

[37] L. Fraiwan, K. Lweesy, N. Khasawneh, M. Fraiwan, H. Wenz, and H. Dickhaus, "Classification of Sleep Stages Using Multi-wavelet Time Frequency Entropy and LDA", en, *Methods of Information in Medicine*, vol. 49, no. 03, pp. 230–237, 2010, ISSN: 0026-1270, 2511-705X. DOI: 10.3414/ME09-01-0054. [Online]. Available: http://www.thieme-connect.de/DOI/DOI?10.3414/ME09-01-0054.

[38] V. C. Helland, A. Gapelyuk, A. Suhrbier, *et al.*, "Investigation of an Automatic Sleep Stage Classification by Means of Multiscorer Hypnogram", en, *Methods of Information in Medicine*, vol. 49, no. 05, pp. 467–472, 2010, ISSN: 0026-1270, 2511-705X. DOI: 10.3414/ME09-02-0052. [Online]. Available: http://www.thieme-connect.de/DOI/DOI?10.3414/ME09-02-0052.

[39] U. R. Acharya, E. C.-P. Chua, K. C. Chua, L. C. Min, and T. Tamura, "ANALYSIS AND AUTOMATIC IDENTIFICATION OF SLEEP STAGES USING HIGHER ORDER SPECTRA", en, *International Journal of Neural Systems*, vol. 20, no. 06, pp. 509–521, Dec. 2010, ISSN: 0129-0657, 1793-6462. DOI: 10.1142/S0129065710002589. [Online]. Available: https://www.worldscientific.com/doi/abs/10.1142/S0129065710002589.

[40] J. R. Shambroom, S. E. Fábregas, and J. Johnstone, "Validation of an automated wireless system to monitor sleep in healthy adults", en, *Journal of Sleep Research*, vol. 21, no. 2, pp. 221–230, Apr. 2012, ISSN: 0962-1105, 1365-2869. DOI: 10.1111/j.1365-2869.2011.00944.x. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/j.1365-2869.2011.00944.x.

[41] S.-T. Pan, C.-E. Kuo, J.-H. Zeng, and S.-F. Liang, "A transition-constrained discrete hidden Markov model for automatic sleep staging", en, *BioMedical Engineering OnLine*, vol. 11, no. 1, p. 52, Dec. 2012, ISSN: 1475-925X. DOI: 10.1186/1475-925X-11-52. [Online]. Available: https://biomedical-engineering-online.biomedcentral.com/articles/10.1186/1475-925X-11-52.

[42] B. Zhang, T. Lei, H. Liu, and H. Cai, "EEG-Based Automatic Sleep Staging Using Ontology and Weighting Feature Analysis", en, *Computational and Mathematical Methods in Medicine*, vol. 2018, pp. 1–16, Sep. 2018, ISSN: 1748-670X, 1748-6718. DOI: 10.1155/2018/6534041. [Online]. Available: https://www.hindawi.com/journals/cmmm/2018/6534041/.

[43] H. Koch, P. Jennum, and J. A. E. Christensen, "Automatic sleep classification using adaptive segmentation reveals an increased number ofrapid eye movement sleep transitions", en, *Journal of Sleep Research*, vol. 28, no. 2, e12780, Apr. 2019, ISSN: 0962-1105, 1365-2869. DOI: 10.1111/jsr.12780. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/jsr.12780.

[44] E. Christensen, A. Abosch, J. A. Thompson, and J. Zylberberg, "Inferring sleep stage from local field potentials recorded in the subthalamic nucleus of Parkinson's patients", en, *Journal of Sleep Research*, vol. 28, no. 4, e12806, Aug. 2019, ISSN: 0962-1105, 1365-2869. DOI: 10.1111/jsr.12806. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/jsr.12806.

[45] T. Zhu, W. Luo, and F. Yu, "Convolution- and Attention-Based Neural Network for Automated Sleep Stage Classification", en, *International Journal of Environmental Research and Public Health*, vol. 17, no. 11, Jun. 2020, ISSN: 1660-4601. DOI: 10.3390/ijerph17114152. [Online]. Available: https://www.mdpi.com/1660-4601/17/11/4152.

[46] H. Phan, K. Mikkelsen, O. Y. Chén, *et al.*, "Personalized automatic sleep staging with single-night data: A pilot study with Kullback–Leibler divergence regularization", *Physiological Measurement*, vol. 41, no. 6, p. 064 004, Jun. 2020, ISSN: 0967-3334, 1361-6579. DOI: 10.1088/1361-6579/ab921e. [Online]. Available: https://iopscience.iop.org/article/10.1088/1361-6579/ab921e.

[47] H. Shen, F. Ran, M. Xu, A. Guez, A. Li, and A. Guo, "An Automatic Sleep Stage Classification Algorithm Using Improved Model Based Essence Features", en, *Sensors*, vol. 20, no. 17, p. 4677, Aug. 2020, ISSN: 1424-8220. DOI: 10.3390/s20174677. [Online]. Available: https://www.mdpi.com/1424-8220/20/17/4677.

[48] P. Fonseca, M. M. van Gilst, M. Radha, *et al.*, "Automatic sleep staging using heart rate variability, body movements, and recurrent neural networks in a sleep disordered population", *Sleep*, vol. 43, no. 9, zsaa048, Sep. 2020, ISSN: 0161-8105. DOI: 10.1093/sleep/zsaa048. [Online]. Available: https://doi.org/10.1093/sleep/zsaa048.

[49] K. M. Gunnarsdottir, C. Gamaldo, R. M. Salas, *et al.*, "A novel sleep stage scoring system: Combining expert-based features with the generalized linear model", en, *Journal of Sleep Research*, vol. 29, no. 5, e12991, Oct. 2020, ISSN: 0962-1105, 1365-2869. DOI: 10.1111/jsr.12991. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/jsr.12991.

[50] R. K. Tripathy, S. K. Ghosh, P. Gajbhiye, and U. R. Acharya, "Development of Automated Sleep Stage Classification System Using Multivariate Projection-Based Fixed Boundary Empirical Wavelet Transform and Entropy Features Extracted from Multichannel EEG Signals", en, *Entropy*, vol. 22, no. 10, p. 1141, Oct. 2020, ISSN: 1099-4300. DOI: 10.3390/e22101141. [Online]. Available: https://www.mdpi.com/1099-4300/22/10/1141.

[51] M. M. Van Gilst, B. M. Wulterkens, P. Fonseca, *et al.*, "Direct application of an ECG-based sleep staging algorithm on reflective photoplethysmography data decreases performance", en, *BMC Research Notes*, vol. 13, no. 1, p. 513, Dec. 2020, ISSN: 1756-0500. DOI: 10.1186/s13104-020-05355-0. [Online]. Available: https://bmcresnotes.biomedcentral.com/articles/10.1186/s13104-020-05355-0.

[52] P. J. Arnal, V. Thorey, E. Debellemaniere, *et al.*, "The Dreem Headband compared to polysomnography for electroencephalographic signal acquisition and sleep staging", en, *Sleep*, vol. 43, no. 11, zsaa097, Nov. 2020, ISSN: 0161-8105, 1550-9109. DOI: 10.1093/sleep/zsaa097. [Online]. Available: https://academic.oup.com/sleep/article/doi/10.1093/sleep/zsaa097/5841249.

[53] H. Korkalainen, J. Aakko, B. Duce, *et al.*, "Deep learning enables sleep staging from photoplethysmogram for patients with suspected sleep apnea", eng, *Sleep*, vol. 43, no. 11, zsaa098, Nov. 2020, ISSN: 1550-9109. DOI: 10.1093/sleep/zsaa098.

[54] T. Zhu, W. Luo, and F. Yu, "Multi-Branch Convolutional Neural Network for Automatic Sleep Stage Classification with Embedded Stage Refinement and Residual Attention Channel Fusion", en, *Sensors*, vol. 20, no. 22, p. 6592, Nov. 2020, ISSN: 1424-8220. DOI: 10.3390/s20226592. [Online]. Available: https://www.mdpi.com/1424-8220/20/22/6592.

[55] X. Zhao and G. Sun, "A Multi-Class Automatic Sleep Staging Method Based on Photoplethysmography Signals", en, *Entropy*, vol. 23, no. 1, p. 116, Jan. 2021, ISSN: 1099-4300. DOI: 10.3390/e23010116. [Online]. Available: https://www.mdpi.com/1099-4300/23/1/116.

[56] A. N. Olesen, P. Jørgen Jennum, E. Mignot, and H. B. D. Sorensen, "Automatic sleep stage classification with deep residual networks in a mixed-cohort setting", en, *Sleep*, vol. 44, no. 1, zsaa161, Jan. 2021, ISSN: 0161-8105, 1550-9109. DOI: 10.1093/sleep/zsaa161. [Online]. Available: https://academic.oup.com/sleep/article/doi/10.1093/sleep/zsaa161/5897250.

[57] M. Fu, Y. Wang, Z. Chen, *et al.*, "Deep Learning in Automatic Sleep Staging With a Single Channel Electroencephalography", *Frontiers in Physiology*, vol. 12, p. 628 502, Mar. 2021, ISSN: 1664-042X. DOI: 10.3389/fphys.2021.628502. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fphys.2021.628502/full.

[58] M. Sharma, J. Tiwari, and U. R. Acharya, "Automatic Sleep-Stage Scoring in Healthy and Sleep Disorder Patients Using Optimal Wavelet Filter Bank Technique with EEG Signals", en, *International Journal of Environmental Research and Public Health*, vol. 18, no. 6, p. 3087, Mar. 2021, ISSN: 1660-4601. DOI: 10.3390/ijerph18063087. [Online]. Available: https://www.mdpi.com/1660-4601/18/6/3087.

[59] A. A. Casciola, S. K. Carlucci, B. A. Kent, *et al.*, "A Deep Learning Strategy for Automatic Sleep Staging Based on Two-Channel EEG Headband Data", en, *Sensors*, vol. 21, no. 10, p. 3316, May 2021, ISSN: 1424-8220. DOI: 10.3390/s21103316. [Online]. Available: https://www.mdpi.com/1424-8220/21/10/3316.

[60] T. Tezuka, D. Kumar, S. Singh, I. Koyanagi, T. Naoi, and M. Sakaguchi, "Real-time, automatic, open-source sleep stage classification system using single EEG for mice", en, *Scientific Reports*, vol. 11, no. 1, p. 11 151, May 2021, ISSN: 2045-2322. DOI: 10.1038/s41598-021-90332-1. [Online]. Available: https://www.nature.com/articles/s41598-021-90332-1.

[61] 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. 573 194, Jul. 2021, ISSN: 1662-453X. DOI: 10.3389/fnins.2021.573194. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnins.2021.573194/full.

[62] L.-X. Feng, X. Li, H.-Y. Wang, *et al.*, "Automatic Sleep Staging Algorithm Based on Time Attention Mechanism", *Frontiers in Human Neuroscience*, vol. 15, p. 692 054, Aug. 2021, ISSN: 1662-5161. DOI: 10.3389/fnhum.2021.692054. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnhum.2021.692054/full.

[63] S. Sun, C. Li, N. Lv, X. Zhang, Z. Yu, and H. Wang, "Attention based convolutional network for automatic sleep stage classification", en, *Biomedical Engineering / Biomedizinische Technik*, vol. 66, no. 4, pp. 335–343, Aug. 2021, ISSN: 1862-278X, 0013-5585. DOI: 10.1515/bmt-2020-0051. [Online]. Available:

https://www.degruyter.com/document/doi/10.1515/bmt-2020-0051/html.

[64] R. Huttunen, T. Leppänen, B. Duce, *et al.*, "Assessment of obstructive sleep apnea-related sleep fragmentation utilizing deep learning-based sleep staging from photoplethysmography", en, *SLEEP*, vol. 44, no. 10, zsab142, Oct. 2021, ISSN: 0161-8105, 1550-9109. DOI: 10.1093/sleep/zsab142. [Online]. Available: https://academic.oup.com/sleep/article/doi/10.1093/sleep/zsab142/6294001.

[65] I. Rechichi, A. Iadarola, M. Zibetti, A. Cicolin, and G. Olmo, "Assessing REM Sleep Behaviour Disorder: From Machine Learning Classification to the Definition of a Continuous Dissociation Index", en, *International Journal of Environmental Research and Public Health*, vol. 19, no. 1, p. 248, Dec. 2021, ISSN: 1660-4601. DOI: 10.3390/ijerph19010248. [Online]. Available: https://www.mdpi.com/1660-4601/19/1/248.

[66] D. Zhao, R. Jiang, M. Feng, *et al.*, "A deep learning algorithm based on 1D CNN-LSTM for automatic sleep staging", *Technology and Health Care*, vol. 30, no. 2, pp. 323–336, Mar. 2022, ISSN: 09287329, 18787401. DOI: 10.3233/THC-212847. [Online]. Available: https://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/THC-212847.

[67] L. Thiesse, L. Staner, P. Bourgin, *et al.*, "Validation of SomnoArt Software, a novel approach of sleep staging, compared with polysomnography in disturbed sleep profiles", en, *SLEEP Advances*, vol. 3, no. 1, zpab019, Jan. 2022, ISSN: 2632-5012. DOI: 10.1093/sleepadvances/zpab019. [Online]. Available: https://academic.oup.com/sleepadvances/article/doi/10.1093/sleepadvances/zpab019/6458489.

[68] D. Kim, J. Lee, Y. Woo, J. Jeong, C. Kim, and D.-K. Kim, "Deep Learning Application to Clinical Decision Support System in Sleep Stage Classification", en, *Journal of Personalized Medicine*, vol. 12, no. 2, p. 136, Jan. 2022, ISSN: 2075-4426. DOI: 10.3390/jpm12020136. [Online]. Available: https://www.mdpi.com/2075-4426/12/2/136.

[69] D. Chylinski, C. Berthomier, E. Lambot, *et al.*, "Variability of sleep stage scoring in late midlife and early old age", en, *Journal of Sleep Research*, vol. 31, no. 1, e13424, Feb. 2022, ISSN: 0962-1105, 1365-2869. DOI: 10.1111/jsr.13424. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/jsr.13424.

[70] S. Zhao, F. Long, X. Wei, X. Ni, H. Wang, and B. Wei, "Evaluation of a Single-Channel EEG-Based Sleep Staging Algorithm", en, *International Journal of Environmental Research and Public Health*, vol. 19, no. 5, p. 2845, Mar. 2022, ISSN: 1660-4601. DOI: 10.3390/ijerph19052845. [Online]. Available: https://www.mdpi.com/1660-4601/19/5/2845.

[71] K. Kwon, S. Kwon, and W.-H. Yeo, "Automatic and Accurate Sleep Stage Classification via a Convolutional Deep Neural Network and Nanomembrane Electrodes", en, *Biosensors*, vol. 12, no. 3, p. 155, Mar. 2022, ISSN: 2079-6374. DOI: 10.3390/bios12030155. [Online]. Available: https://www.mdpi.com/2079-6374/12/3/155.

[72] Y. Hei, T. Yuan, Z. Fan, B. Yang, and J. Hu, "Sleep staging classification based on a new parallel fusion method of multiple sources signals", *Physiological Measurement*, vol. 43, no. 4, p. 045 003, Apr. 2022, ISSN: 0967-3334, 1361-6579. DOI: 10.1088/1361-6579/ac647b. [Online]. Available: https://iopscience.iop.org/article/10.1088/1361-6579/ac647b.

[73] N. Von Ellenrieder, L. Peter-Derex, J. Gotman, and B. Frauscher, "SleepSEEG: Automatic sleep scoring using intracranial EEG recordings only", *Journal of Neural Engineering*, vol. 19, no. 2, p. 026 057, Apr. 2022, ISSN: 1741-2560, 1741-2552. DOI: 10.1088/1741-2552/ac6829. [Online]. Available: https://iopscience.iop.org/article/10.1088/1741-2552/ac6829.

[74] M. Sharma, A. Yadav, J. Tiwari, M. Karabatak, O. Yildirim, and U. R. Acharya, "An Automated Wavelet-Based Sleep Scoring Model Using EEG, EMG, and EOG Signals with More Than 8000 Subjects", en, *International Journal of Environmental Research and Public Health*, vol. 19, no. 12, p. 7176, Jun. 2022, ISSN: 1660-4601. DOI: 10.3390/ijerph19127176. [Online]. Available: https://www.mdpi.com/1660-4601/19/12/7176.

[75] C. Liu, Y. Yin, Y. Sun, and O. K. Ersoy, "Multi-scale ResNet and BiGRU automatic sleep staging based on attention mechanism", en, *PLOS ONE*, vol. 17, no. 6, S. V E, Ed., e0269500, Jun. 2022, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0269500. [Online]. Available: https://dx.plos.org/10.1371/journal.pone.0269500.

[76] Y. Li, Z. Xu, Y. Zhang, Z. Cao, and H. Chen, "Automatic sleep stage classification based on a two-channel electrooculogram and one-channel electromyogram", *Physiological Measurement*, vol. 43, no. 7, 07NT02, Jul. 2022, ISSN: 0967-3334, 1361-6579. DOI: 10.1088/1361-6579/ac6bdb. [Online]. Available: https://iopscience.iop.org/article/10.1088/1361-6579/ac6bdb.

[77] G. Zou, J. Liu, Q. Zou, and J.-H. Gao, "A-PASS: An automated pipeline to analyze simultaneously acquired EEG-fMRI data for studying brain activities during sleep", *Journal of Neural Engineering*, vol. 19, no. 4, p. 046 031, Aug. 2022, ISSN: 1741-2560, 1741-2552. DOI: 10.1088/1741-2552/ac83f2. [Online]. Available: https://iopscience.iop.org/article/10.1088/1741-2552/ac83f2.

[78] A.-M. Tăuțan, A. C. Rossi, and B. Ionescu, "Automatic sleep scoring with LSTM networks: Impact of time granularity and input signals", en, *Biomedical Engineering / Biomedizinische Technik*, vol. 67, no. 4, pp. 267–281, Aug. 2022, ISSN: 0013-5585, 1862-278X. DOI: 10.1515/bmt-2021-0408. [Online]. Available: https://www.degruyter.com/document/doi/10.1515/bmt-2021-0408/html.

[79] X. Lv, J. Li, and Q. Xu, "A Multilevel Temporal Context Network for Sleep Stage Classification", en, *Computational Intelligence and Neuroscience*, vol. 2022, H. Cecotti, Ed., pp. 1–15, Sep. 2022, ISSN: 1687-5273, 1687-5265. DOI: 10.1155/2022/6104736. [Online]. Available: https://www.hindawi.com/journals/cin/2022/6104736/.

[80] H. ElMoaqet, M. Eid, M. Ryalat, and T. Penzel, "A Deep Transfer Learning Framework for Sleep Stage Classification with Single-Channel EEG Signals", en, *Sensors*, vol. 22, no. 22, p. 8826, Nov. 2022, ISSN: 1424-8220. DOI: 10.3390/s22228826. [Online]. Available: https://www.mdpi.com/1424-8220/22/22/8826.

[81] Y. You, X. Guo, X. Zhong, and Z. Yang, "A Few-Shot Learning-Based EEG and Stage Transition Sequence Generator for Improving Sleep Staging Performance", en, *Biomedicines*, vol. 10, no. 12, p. 3006, Nov. 2022, ISSN: 2227-9059. DOI: 10.3390/biomedicines10123006. [Online]. Available: https://www.mdpi.com/2227-9059/10/12/3006.

[82] R. Yu, Z. Zhou, S. Wu, X. Gao, and G. Bin, "MRASleepNet: A multi-resolution attention network for sleep stage classification using single-channel EEG", *Journal of Neural Engineering*, vol. 19, no. 6, p. 066 025, Dec. 2022, ISSN: 1741-2560, 1741-2552. DOI: 10.1088/1741-2552/aca2de. [Online]. Available: https://iopscience.iop.org/article/10.1088/1741-2552/aca2de.

[83] Y. R. Tabar, K. B. Mikkelsen, N. Shenton, *et al.*, "At-home sleep monitoring using generic ear-EEG", *Frontiers in Neuroscience*, vol. 17, p. 987 578, Feb. 2023, ISSN: 1662-453X. DOI: 10.3389/fnins.2023.987578. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnins.2023.987578/full.

[84] M. Cesari, K. Egger, A. Stefani, *et al.*, "Differentiation of central disorders of hypersomnolence with manual and artificial-intelligence-derived polysomnographic measures", en, *Sleep*, vol. 46, no. 2, zsac288, Feb. 2023, ISSN: 0161-8105, 1550-9109. DOI: 10.1093/sleep/zsac288. [Online]. Available: https://academic.oup.com/sleep/article/doi/10.1093/sleep/zsac288/6862127.

[85] S. Haghayegh, K. Hu, K. Stone, S. Redline, and E. Schernhammer, "Automated Sleep Stages Classification Using Convolutional Neural Network From Raw and Time-Frequency Electroencephalogram Signals: Systematic Evaluation Study", en, *Journal of Medical Internet Research*, vol. 25, e40211, Feb. 2023, ISSN: 1438-8871. DOI: 10.2196/40211. [Online]. Available: https://www.jmir.org/2023/1/e40211.

[86] P. Topalidis, D. P. J. Heib, S. Baron, E.-S. Eigl, A. Hinterberger, and M. Schabus, "The Virtual Sleep Lab—A Novel Method for Accurate Four-Class Sleep Staging Using Heart-Rate Variability from Low-Cost Wearables", en, *Sensors*, vol. 23, no. 5, p. 2390, Feb. 2023, ISSN: 1424-8220. DOI: 10.3390/s23052390. [Online]. Available: https://www.mdpi.com/1424-8220/23/5/2390.

[87] Y.-H. Cheng, M. Lech, and R. H. Wilkinson, "Simultaneous Sleep Stage and Sleep Disorder Detection from Multimodal Sensors Using Deep Learning", en, *Sensors*, vol. 23, no. 7, p. 3468, Mar. 2023, ISSN: 1424-8220. DOI: 10.3390/s23073468. [Online]. Available: https://www.mdpi.com/1424-8220/23/7/3468.

[88] P. Somaskandhan, T. Leppänen, P. I. Terrill, *et al.*, "Deep learning-based algorithm accurately classifies sleep stages in preadolescent children with sleep-disordered breathing symptoms and age-matched controls", *Frontiers in Neurology*, vol. 14, p. 1 162 998, Apr. 2023, ISSN: 1664-2295. DOI: 10.3389/fneur.2023.1162998. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fneur.2023.1162998/full.

[89] H. Zhu, C. Fu, F. Shu, H. Yu, C. Chen, and W. Chen, "The Effect of Coupled Electroencephalography Signals in Electrooculography Signals on Sleep Staging Based on Deep Learning Methods", en, *Bioengineering*, vol. 10, no. 5, p. 573, May 2023, ISSN: 2306-5354. DOI: 10.3390/bioengineering10050573. [Online]. Available: https://www.mdpi.com/2306-5354/10/5/573.

[90] T. I. Toma and S. Choi, "An End-to-End Multi-Channel Convolutional Bi-LSTM Network for Automatic Sleep Stage Detection", en, *Sensors*, vol. 23, no. 10, p. 4950, May 2023, ISSN: 1424-8220. DOI: 10.3390/s23104950. [Online]. Available: https://www.mdpi.com/1424-8220/23/10/4950.

[91] P. Fonseca, M. Ross, A. Cerny, *et al.*, "A computationally efficient algorithm for wearable sleep staging in clinical populations", en, *Scientific Reports*, vol. 13, no. 1, p. 9182, Jun. 2023, ISSN: 2045-2322. DOI: 10.1038/s41598-023-36444-2. [Online]. Available: https://www.nature.com/articles/s41598-023-36444-2.

[92] C. Kang, S. An, H. J. Kim, *et al.*, "Age-integrated artificial intelligence framework for sleep stage classification and obstructive sleep apnea screening", *Frontiers in Neuroscience*, vol. 17, p. 1 059 186, Jun. 2023, ISSN: 1662-453X. DOI: 10.3389/fnins.2023.1059186. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnins.2023.1059186/full.

[93] D. Zhang, J. Sun, Y. She, *et al.*, "A two-branch trade-off neural network for balanced scoring sleep stages on multiple cohorts", *Frontiers in Neuroscience*, vol. 17, p. 1 176 551, Jun. 2023, ISSN: 1662-453X. DOI: 10.3389/fnins.2023.1176551. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnins.2023.1176551/full.

[94] C. Yang, B. Li, Y. Li, Y. He, and Y. Zhang, "LWSleepNet: A lightweight attention-based deep learning model for sleep staging with singlechannel EEG", en, *DIGITAL HEALTH*, vol. 9, p. 20 552 076 231 188 206, Jan. 2023, ISSN: 2055-2076, 2055-2076. DOI: 10.1177/20552076231188206. [Online]. Available: http://journals.sagepub.com/doi/10.1177/20552076231188206.

[95] N. Van Den Broek, F. Van Meulen, M. Ross, *et al.*, "Automated sleep staging in people with intellectual disabilities using heart rate and respiration variability", en, *Journal of Intellectual Disability Research*, vol. 67, no. 8, pp. 720–733, Aug. 2023, ISSN: 0964-2633, 1365-2788. DOI: 10.1111/jir.13060. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/jir.13060.

[96] M. Lee, H.-G. Kwak, H.-J. Kim, D.-O. Won, and S.-W. Lee, "SeriesSleepNet: An EEG time series model with partial data augmentation for automatic sleep stage scoring", *Frontiers in Physiology*, vol. 14, p. 1 188 678, Aug. 2023, ISSN: 1664-042X. DOI: 10.3389/fphys.2023.1188678. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fphys.2023.1188678/full.

[97] W. Li and J. Gao, "Automatic sleep staging by a hybrid model based on deep 1D-ResNet-SE and LSTM with single-channel raw EEG signals", en, *PeerJ Computer Science*, vol. 9, e1561, Sep. 2023, ISSN: 2376-5992. DOI: 10.7717/peerj-cs.1561. [Online]. Available: https://peerj.com/articles/cs-1561.

[98] S. Hangaragi, N. Nizampatnam, D. Kaliyaperumal, and T. Özer, "An evolutionary model for sleep quality analytics using fuzzy system", en, *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine*, vol. 237, no. 10, pp. 1215–1227, Oct. 2023, ISSN: 0954-4119, 2041-3033. DOI: 10.1177/09544119231195177. [Online]. Available: http://journals.sagepub.com/doi/10.1177/09544119231195177.

[99] W. Wang, J. Li, Y. Fang, Y. Zheng, and F. You, "An effective hybrid feature selection using entropy weight method for automatic sleep staging", *Physiological Measurement*, vol. 44, no. 10, p. 105 008, Oct. 2023, ISSN: 0967-3334, 1361-6579. DOI: 10.1088/1361-6579/acff35. [Online]. Available: https://iopscience.iop.org/article/10.1088/1361-6579/acff35.

[100] J. Lee, H. C. Kim, Y. J. Lee, and S. Lee, "Development of generalizable automatic sleep staging using heart rate and movement based on large databases", en, *Biomedical Engineering Letters*, vol. 13, no. 4, pp. 649–658, Nov. 2023, ISSN: 2093-9868, 2093-985X. DOI: 10.1007/s13534-023-00288-6. [Online]. Available: https://link.springer.com/10.1007/s13534-023-00288-6.

[101] P. I. Topalidis, S. Baron, D. P. J. Heib, E.-S. Eigl, A. Hinterberger, and M. Schabus, "From Pulses to Sleep Stages: Towards Optimized Sleep Classification Using Heart-Rate Variability", en, *Sensors*, vol. 23, no. 22, p. 9077, Nov. 2023, ISSN: 1424-8220. DOI: 10.3390/s23229077. [Online]. Available: https://www.mdpi.com/1424-8220/23/22/9077.

[102] J. Jeong, W. Yoon, J.-G. Lee, *et al.*, "Standardized image-based polysomnography database and deep learning algorithm for sleep-stage classification", en, *SLEEP*, vol. 46, no. 12, zsad242, Dec. 2023, ISSN: 0161-8105, 1550-9109. DOI: 10.1093/sleep/zsad242. [Online]. Available: https://academic.oup.com/sleep/article/doi/10.1093/sleep/zsad242/7273094.

[103] I. S. Masad, A. Alqudah, and S. Qazan, "Automatic classification of sleep stages using EEG signals and convolutional neural networks", en, *PLOS ONE*, vol. 19, no. 1, A. P. P. Abdul Majeed, Ed., e0297582, Jan. 2024, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0297582. [Online]. Available: https://dx.plos.org/10.1371/journal.pone.0297582.

[104] W. Pei, Y. Li, P. Wen, F. Yang, and X. Ji, "An automatic method using MFCC features for sleep stage classification", en, *Brain Informatics*, vol. 11, no. 1, p. 6, Dec. 2024, ISSN: 2198-4018, 2198-4026. DOI: 10.1186/s40708-024-00219-w. [Online]. Available: https://braininformatics.springeropen.com/articles/10.1186/s40708-024-00219-w.

[105] H. Zhu, Y. Xu, Y. Wu, *et al.*, "A Sequential End-to-End Neonatal Sleep Staging Model with Squeeze and Excitation Blocks and Sequential Multi-Scale Convolution Neural Networks", en, *International Journal of Neural Systems*, vol. 34, no. 03, p. 2 450 013, Mar. 2024, ISSN: 0129-0657, 1793-6462. DOI: 10.1142/S0129065724500138. [Online]. Available: https://www.worldscientific.com/doi/10.1142/S0129065724500138.

[106] H. Han, M. J. Seong, J. Hyeon, E. Joo, and J. Oh, "Classification and automatic scoring of arousal intensity during sleep stages using machine learning", en, *Scientific Reports*, vol. 14, no. 1, p. 5983, Mar. 2024, ISSN: 2045-2322. DOI: 10.1038/s41598-023-50653-9. [Online]. Available: https://www.nature.com/articles/s41598-023-50653-9.

[107] L. Cerina, S. Overeem, G. B. Papini, *et al.*, "A sleep stage estimation algorithm based on cardiorespiratory signals derived from a suprasternal pressure sensor", en, *Journal of Sleep Research*, vol. 33, no. 2, e14015, Apr. 2024, ISSN: 0962-1105, 1365-2869. DOI: 10.1111/jsr.14015. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/jsr.14015.

[108] 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*, pp. 1–1, 2022, arXiv:2105.11043 [cs, eess], ISSN: 0018-9294, 1558-2531. DOI: 10.1109/TBME.2022.3147187. [Online]. Available: http://arxiv.org/abs/2105.11043.

[109] B. Thordarson, G. Jouan, E. Hardarson, *et al.*, *An Optimized Framework for Processing Large-scale Polysomnographic Data Incorporating Expert Human Oversight*, Apr. 2024. DOI: 10.13140/RG.2.2.15631.19361.

[110] M. Rusanen, G. Jouan, R. Huttunen, *et al.*, *aSAGA: Automatic Sleep Analysis with Gray Areas*, arXiv:2310.02032 [cs], Oct. 2023. DOI: 10.48550/arXiv.2310.02032. [Online]. Available: http://arxiv.org/abs/2310.02032.

[111] G. Jouan, E. S. Arnardottir, A. S. Islind, and M. Óskarsdóttir, "An algorithmic approach to identification of gray areas: Analysis of sleep scoring expert ensemble non agreement areas using a multinomial mixture model", *European Journal of Operational Research*, Oct. 2023, ISSN: 0377-2217. DOI: 10.1016/j.ejor.2023.09.039. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0377221723007567.

[112] E. Mosqueira-Rey, E. Hernández-Pereira, D. Alonso-Ríos, J. Bobes-Bascarán, and Á. Fernández-Leal, "Human-in-the-loop machine learning: A state of the art", en, *Artificial Intelligence Review*, vol. 56, no. 4, pp. 3005–3054, Apr. 2023, ISSN: 1573-7462. DOI: 10.1007/s10462-022-10246-w. [Online]. Available: https://doi.org/10.1007/s10462-022-10246-w.

[113] R. K. Lindsay, B. G. Buchanan, E. A. Feigenbaum, and J. Lederberg, "DENDRAL: A case study of the first expert system for scientific hypothesis formation", *Artificial Intelligence*, vol. 61, no. 2, pp. 209–261, Jun. 1993, ISSN: 0004-3702. DOI: 10.1016/0004-3702(93)90068-M. [Online]. Available: https://www.sciencedirect.com/science/article/pii/000437029390068M.

[114] E. H. Shortliffe, R. Davis, S. G. Axline, B. G. Buchanan, C. C. Green, and S. N. Cohen, "Computer-based consultations in clinical therapeutics: Explanation and rule acquisition capabilities of the MYCIN system", eng, *Computers and Biomedical Research, an International Journal*, vol. 8, no. 4, pp. 303–320, Aug. 1975, ISSN: 0010-4809. DOI: 10.1016/0010-4809(75)90009-9.

[115] A. Bottrighi, F. Grosso, M. Ghiglione, *et al.*, "A Symbolic AI Approach to Medical Training", *Journal of Medical Systems*, vol. 49, no. 1, p. 2, 2025, ISSN: 0148-5598. DOI: 10.1007/s10916-024-02139-y. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC11717836/.

[116] B. Reddy and R. Fields, "From past to present: A comprehensive technical review of rule-based expert systems from 1980 – 2021", in *Proceedings of the 2022 ACM Southeast Conference*, ser. ACMSE '22, New York, NY, USA: Association for Computing Machinery, May 2022, pp. 167–172, ISBN: 978-1-4503-8697-5. DOI: 10.1145/3476883.3520211. [Online]. Available: https://dl.acm.org/doi/10.1145/3476883.3520211.

[117] G. Hripcsak, "Writing Arden Syntax Medical Logic Modules", *Computers in Biology and Medicine*, vol. 24, no. 5, pp. 331–363, 1994. DOI: 10.1016/0010-4825(94)90002-7.

[118] M. Samwald, K. Fehre, J. de Bruin, and K.-P. Adlassnig, "The Arden Syntax standard for clinical decision support: Experiences and directions", *Journal of Biomedical Informatics*, vol. 45, no. 4, pp. 711–718, 2012. DOI: 10.1016/j.jbi.2012.02.001.

[119] G. Hripcsak, J. D. Duke, N. H. Shah, *et al.*, "Observational health data sciences and informatics (OHDSI): Opportunities for observational researchers", in *Studies in Health Technology and Informatics*, vol. 216, 2015, pp. 574–578.

[120] D. Bender and K. Sartipi, "HL7 FHIR: An agile and RESTful approach to healthcare information exchange", in *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems*, 2013, pp. 326–331. DOI: 10.1109/CBMS.2013.6627810.

[121] J. C. Mandel, D. A. Kreda, K. D. Mandl, I. S. Kohane, and R. B. Ramoni, "SMART on FHIR: A standards-based, interoperable apps platform for electronic health records", *Journal of the American Medical Informatics Association*, vol. 23, no. 5, pp. 899–908, 2016. DOI: 10.1093/jamia/ocv189.

[122] HL7 International, *Clinical Quality Language (CQL) specification, release 1 STU 5*, https://cql.hl7.org/, Accessed April 2026, 2023.

[123] HL7 International, *CDS Hooks 2.0 implementation guide*, https://cds-hooks.hl7.org/2.0/, Accessed April 2026, 2022.

[124] P. Taber, C. Radloff, G. Del Fiol, C. Staes, and K. Kawamoto, "New standards for clinical decision support: A survey of the state of implementation", *Yearbook of Medical Informatics*, vol. 30, no. 1, pp. 159–171, 2021. DOI: 10.1055/s-0041-1726502.

[125] H. R. Strasberg, B. Rhodes, G. Del Fiol, R. A. Jenders, P. J. Haug, and K. Kawamoto, "Contemporary clinical decision support standards using Health Level Seven International Fast Healthcare Interoperability Resources", *Journal of the American Medical Informatics Association*, vol. 28, no. 8, pp. 1796–1806, 2021. DOI: 10.1093/jamia/ocab070.

[126] K. R. Thórisson and D. Kremelberg, "Understanding and Common Sense: Two Sides of the Same Coin?", en, in *Artificial General Intelligence*, T. Everitt, B. Goertzel, and A. Potapov, Eds., Cham: Springer International Publishing, 2017, pp. 201–211, ISBN: 978-3-319-63703-7. DOI: 10.1007/978-3-319-63703-7_19.

[127] B. G. Buchanan and R. G. Smith, "Fundamentals of Expert Systems", en, *Annual Review of Computer Science*, vol. 3, no. Volume 3, 1988, pp. 23–58, Jun. 1988, ISSN: 8756-7016. DOI: 10.1146/annurev.cs.03.060188.000323. [Online]. Available: https://www.annualreviews.org/content/journals/10.1146/annurev.cs.03.060188.000323.

[128] K. Metaxiotis and J. Psarras, "Expert systems in business: Applications and future directions for the operations researcher", en, *Industrial Management & Data Systems*, vol. 103, no. 5, pp. 361–368, Jul. 2003, ISSN: 0263-5577. DOI: 10.1108/02635570310477.

[Online]. Available: http://www.emerald.com/imds/article/103/5/361-368/179852.

[129] K. P. Murphy, *Machine learning: a probabilistic perspective* (Adaptive computation and machine learning series). Cambridge, MA: MIT Press, 2012, ISBN: 978-0-262-01802-9.

[130] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI", *Information Fusion*, vol. 58, pp. 82–115, Jun. 2020, ISSN: 1566-2535. DOI: 10.1016/j.inffus.2019.12.012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253519308103.

[131] Z. C. Lipton, "The mythos of model interpretability", *Commun. ACM*, vol. 61, no. 10, pp. 36–43, Sep. 2018, ISSN: 0001-0782. DOI: 10.1145/3233231. [Online]. Available: https://dl.acm.org/doi/10.1145/3233231.

[132] D. Castelvecchi, "Can we open the black box of AI?", en, *Nature News*, vol. 538, no. 7623, p. 20, Oct. 2016, Cg_type: Nature News. DOI: 10.1038/538020a. [Online]. Available: http://www.nature.com/news/can-we-open-the-black-box-of-ai-1.20731.

[133] G. Vilone and L. Longo, *Explainable Artificial Intelligence: A Systematic Review*, arXiv:2006.00093 [cs], Oct. 2020. DOI: 10.48550/arXiv.2006.00093. [Online]. Available: http://arxiv.org/abs/2006.00093.

[134] J. C. Pitt, *Theories of Explanation*. Cary, UNITED STATES: Oxford University Press USA - OSO, 1988, ISBN: 978-0-19-802166-7.

[135] D.-H. Ruben, *Explaining Explanation*. London, UNITED KINGDOM: Taylor & Francis Group, 2012, ISBN: 978-1-315-63473-9.

[136] T. Hulsen, "Explainable Artificial Intelligence (XAI): Concepts and Challenges in Healthcare", en, *AI*, vol. 4, no. 3, pp. 652–666, Sep. 2023, Number: 3, ISSN: 2673-2688. DOI: 10.3390/ai4030034. [Online]. Available: https://www.mdpi.com/2673-2688/4/3/34.

[137] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, "The false hope of current approaches to explainable artificial intelligence in health care", English, *The Lancet Digital Health*, vol. 3, no. 11, e745–e750, Nov. 2021, ISSN: 2589-7500. DOI: 10.1016/S2589-7500(21)00208-9. [Online]. Available: https://www.thelancet.com/journals/landig/article/PIIS2589-7500(21)00208-9/fulltext.

[138] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead", *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, May 2019, ISSN: 2522-5839. DOI: 10.1038/s42256-019-0048-x. [Online]. Available: https://doi.org/10.1038/s42256-019-0048-x.

[139] B. Babic and I. G. Cohen, *The Algorithmic Explainability 'Bait and Switch'*, en, SSRN Scholarly Paper, Rochester, NY, Aug. 2023. [Online]. Available: https://papers.ssrn.com/abstract=4541487.

[140] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences", *Artificial Intelligence*, vol. 267, pp. 1–38, Feb. 2019, ISSN: 0004-3702. DOI: 10.1016/j.artint.2018.07.007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0004370218305988.

[141] H. L. A. Hart and P. A. Bulloch, *The concept of law* (Clarendon law series), eng, 2. ed., repr. Oxford [u.a]: Clarendon Press, 1998, ISBN: 978-0-19-876123-5.

[142] *VEHICLE Definition & Meaning - Merriam-Webster*. [Online]. Available: https://www.merriam-webster.com/dictionary/vehicle.

[143] J. Vervaeke, T. P. Lillicrap, and B. A. Richards, "Relevance Realization and the Emerging Framework in Cognitive Science", *Journal of Logic and Computation*, vol. 22, no. 1, pp. 79–99, Feb. 2012, ISSN: 1465-363X. DOI: 10.1093/logcom/exp067. [Online]. Available: https://ieeexplore.ieee.org/document/8132809.

[144] A. d. Garcez and L. C. Lamb, "Neurosymbolic AI: The 3rd wave", en, *Artificial Intelligence Review*, vol. 56, no. 11, pp. 12 387–12 406, Nov. 2023, ISSN: 1573-7462. DOI: 10.1007/s10462-023-10448-w. [Online]. Available: https://doi.org/10.1007/s10462-023-10448-w.

[145] S. Kierner, J. Kucharski, and Z. Kierner, "Taxonomy of hybrid architectures involving rule-based reasoning and machine learning in clinical decision systems: A scoping review", *Journal of Biomedical Informatics*, vol. 144, p. 104 428, Aug. 2023, ISSN: 1532-0464. DOI: 10.1016/j.jbi.2023.104428. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1532046423001491.

[146] H. A. Kautz, "The third AI summer: AAAI Robert S. Engelmore Memorial Lecture", en, *AI Magazine*, vol. 43, no. 1, pp. 105–125, 2022, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/aaai.12036, ISSN: 2371-9621. DOI: 10.1002/aaai.12036. [Online]. Available:

https://onlinelibrary.wiley.com/doi/abs/10.1002/aaai.12036.

[147] G. A. Prenosil, T. K. Weitzel, S. C. Bello, *et al.*, "Neuro-symbolic AI for auditable cognitive information extraction from medical reports", en, *Communications Medicine*, vol. 5, no. 1, p. 491, Nov. 2025, ISSN: 2730-664X. DOI: 10.1038/s43856-025-01194-x. [Online]. Available: https://www.nature.com/articles/s43856-025-01194-x.

[148] J. Kauffman, E. Holmes, A. Vaid, *et al.*, "InfEHR: Clinical phenotype resolution through deep geometric learning on electronic health records", en, *Nature Communications*, vol. 16, no. 1, p. 8475, Sep. 2025, ISSN: 2041-1723. DOI: 10.1038/s41467-025-63366-6. [Online]. Available: https://www.nature.com/articles/s41467-025-63366-6.

[149] F. Ma, J. Gao, Q. Suo, Q. You, J. Zhou, and A. Zhang, "Risk Prediction on Electronic Health Records with Prior Medical Knowledge", in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ser. KDD '18, New York, NY, USA: Association for Computing Machinery, Jul. 2018, pp. 1910–1919, ISBN: 978-1-4503-5552-0. DOI: 10.1145/3219819.3220020. [Online]. Available: https://dl.acm.org/doi/10.1145/3219819.3220020.

[150] W. G. La Cava, P. C. Lee, I. Ajmal, *et al.*, "A flexible symbolic regression method for constructing interpretable clinical prediction models", en, *npj Digital Medicine*, vol. 6, no. 1, p. 107, Jun. 2023, ISSN: 2398-6352. DOI: 10.1038/s41746-023-00833-8. [Online]. Available: https://www.nature.com/articles/s41746-023-00833-8.

[151] J. Dong and J. Zhong, "Recent Advances in Symbolic Regression", en, *ACM Computing Surveys*, vol. 57, no. 11, pp. 1–37, Nov. 2025, ISSN: 0360-0300, 1557-7341. DOI: 10.1145/3735634. [Online]. Available: https://dl.acm.org/doi/10.1145/3735634.

[152] J. Corny, A. Rajkumar, O. Martin, *et al.*, "A machine learning–based clinical decision support system to identify prescriptions with a high risk of medication error", *Journal of the American Medical Informatics Association*, vol. 27, no. 11, pp. 1688–1694, Nov. 2020, ISSN: 1527-974X. DOI: 10.1093/jamia/ocaa154. [Online]. Available: https://doi.org/10.1093/jamia/ocaa154.

[153] S. Kierner, P. Kierner, and J. Kucharski, "Combining machine learning models and rule engines in clinical decision systems: Exploring optimal aggregation methods for vaccine hesitancy prediction", *Computers in Biology and Medicine*, vol. 188, p. 109 749, Apr. 2025, ISSN: 0010-4825. DOI: 10.1016/j.compbiomed.2025.109749. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S001048252500099X.

[154] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design science in information systems research", *MIS Q.*, vol. 28, no. 1, pp. 75–105, Mar. 2004, ISSN: 0276-7783.

[155] J. vom Brocke, A. Hevner, and A. Maedche, "Introduction to Design Science Research", en, in *Design Science Research. Cases*, J. vom Brocke, A. Hevner, and A. Maedche, Eds., Cham: Springer International Publishing, 2020, pp. 1–13, ISBN: 978-3-030-46781-4. DOI: 10.1007/978-3-030-46781-4_1. [Online]. Available: https://doi.org/10.1007/978-3-030-46781-4_1.

[156] K. Peffers, T. Tuunanen, M. Rothenberger, and S. Chatterjee, "A Design Science Research Methodology for Information Systems Research", *J. Manage. Inf. Syst.*, vol. 24, no. 3, pp. 45–77, Dec. 2007, ISSN: 0742-1222. DOI: 10.2753/MIS0742-1222240302. [Online]. Available: https://doi.org/10.2753/MIS0742-1222240302.

[157] G.-Q. Zhang, L. Cui, R. Mueller, *et al.*, "The National Sleep Research Resource: Towards a sleep data commons", eng, *Journal of the American Medical Informatics Association: JAMIA*, vol. 25, no. 10, pp. 1351–1358, Oct. 2018, ISSN: 1527-974X. DOI: 10.1093/jamia/ocy064.

[158] S. F. Quan, B. V. Howard, C. Iber, *et al.*, "The Sleep Heart Health Study: Design, rationale, and methods", eng, *Sleep*, vol. 20, no. 12, pp. 1077–1085, Dec. 1997, ISSN: 0161-8105.

[159] M. S. Mourtazaev, B. Kemp, A. H. Zwinderman, and H. A. Kamphuisen, "Age and gender affect different characteristics of slow waves in the sleep EEG", eng, *Sleep*, vol. 18, no. 7, pp. 557–564, Sep. 1995, ISSN: 0161-8105. DOI: 10.1093/sleep/18.7.557.

[160] E. S. Arnardottir, A. S. Islind, M. Óskarsdóttir, *et al.*, "The Sleep Revolution project: The concept and objectives", en, *Journal of Sleep Research*, vol. 31, no. 4, e13630, 2022, ISSN: 1365-2869. DOI: 10.1111/jsr.13630. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/jsr.13630.

[161] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy", *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[162] D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: An overview of methods, challenges, and prospects", *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1449–1477, 2015.

[163] G. R. Ghosal and R. Abbasi-Asl, "Multi-Modal Prototype Learning for Interpretable Multivariable Time Series Classification", *arXiv.org*, 2021.

[164] S. El-Sappagh, T. Abuhmed, S. Riazul Islam, and K. S. Kwak, "Multimodal multitask deep learning model for Alzheimer's disease progression detection based on time series data", *Neurocomputing*, vol. 412, pp. 197–215, Oct. 2020, ISSN: 0925-2312. DOI: 10.1016/j.neucom.2020.05.087. [Online]. Available: http://dx.doi.org/10.1016/j.neucom.2020.05.087.

[165] S. Li, Y. Li, and Y. Fu, "Multi-View Time Series Classification", in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, ACM, Oct. 2016, pp. 989–998. DOI: 10.1145/2983323.2983780. [Online]. Available: http://dx.doi.org/10.1145/2983323.2983780.

[166] M. S. K. Inan, K. Liao, H. Shen, P. Jayaraman, F. Montori, and D. Georgakopoulos, "Deepmetaiot: A Multimodal Deep Learning Framework Harnessing Metadata for IoT Sensor Data Classification", *IEEE Internet of Things Journal*, 2025. DOI: 10.1109/JIOT.2025.3595556.

[167] 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, 2018.

[168] S. K. Karmaker, M. M. Hassan, M. J. Smith, L. Xu, C. Zhai, and K. Veeramachaneni, "Automl to date and beyond: Challenges and opportunities", *Acm computing surveys (csur)*, vol. 54, no. 8, pp. 1–36, 2021.

[169] B. Zoph and Q. V. Le, "Neural architecture search with reinforcement learning", *arXiv preprint arXiv:1611.01578*, 2016.

[170] M.-A. Zöller and M. F. Huber, "Benchmark and survey of automated machine learning frameworks", *Journal of artificial intelligence research*, vol. 70, pp. 409–472, 2021.

[171] P. Ren, Y. Xiao, X. Chang, *et al.*, "A comprehensive survey of neural architecture search: Challenges and solutions", *ACM Computing Surveys (CSUR)*, vol. 54, no. 4, pp. 1–34, 2021.

[172] M. Tavakoli, R. Chandra, F. Tian, and C. Bravo, "Multi-modal deep learning for credit rating prediction using text and numerical data streams", *Applied Soft Computing*, vol. 171, p. 112 771, 2025.

[173] S. T. Bastos, R. d. C. Silva, and D. F. Silva, "Multimodal Fusion Strategies for Multivariate Time Series Classification", in *Anais do XXII Encontro Nacional de Inteligência Artificial e Computacional (ENIAC 2025)*, Sociedade Brasileira de Computação - SBC, Sep. 2025, pp. 1209–1220. DOI: 10.5753/eniac.2025.14459. [Online]. Available: http://dx.doi.org/10.5753/eniac.2025.14459.

[174] N. Erickson, J. Mueller, A. Shirkov, *et al.*, "Autogluon-tabular: Robust and accurate automl for structured data", *arXiv preprint arXiv:2003.06505*, 2020.

[175] C. MacKinnon and R. Atkinson, "Designing a New Search Space for Multivariate Time-Series Neural Architecture Search", *AALTD@ECML/PKDD*, 2023. DOI: 10.1007/978-3-031-49896-1_13.

[176] N. Miguel Morante, R. Z. Iván, D. García Prieto, R. Duque Medina, C. Palazuelos Calderón, A. Casas García-Minguillán, *et al.*, "Integrating large language models into automated machine learning: A human-centric approach", *KDIR 2025 - 17th International Conference on Knowledge Discovery and Information Retrieval*, 2025.

[177] P. Trirat, W. Jeong, and S. J. Hwang, "Automl-agent: A multi-agent llm framework for full-pipeline automl", *Proceedings of the 42 nd International Conference on Machine Learning*, 2024.

[178] E. LeDell, S. Poirier, *et al.*, "H2o automl: Scalable automatic machine learning", in *Proceedings of the AutoML Workshop at ICML*, vol. 2020, 2020, p. 24.

[179] M. Feurer, K. Eggensperger, S. Falkner, M. Lindauer, and F. Hutter, "Auto-sklearn 2.0: Hands-free automl via meta-learning", *Journal of Machine Learning Research*, vol. 23, no. 261, pp. 1–61, 2022.

[180] A. P. Ruiz, M. Flynn, J. Large, M. Middlehurst, and A. Bagnall, "The great multivariate time series classification bake off: A review and experimental evaluation of recent algorithmic advances", en, *Data Mining and Knowledge Discovery*, vol. 35, no. 2, pp. 401–449, Mar. 2021, ISSN: 1573-756X. DOI: 10.1007/s10618-020-00727-3. [Online]. Available: https://doi.org/10.1007/s10618-020-00727-3.

[181] B. Kemp, A. H. Zwinderman, B. Tuk, H. A. Kamphuisen, and J. J. Oberye, "Analysis of a sleep-dependent neuronal feedback loop: The slow-wave microcontinuity of the eeg", *IEEE Transactions on Biomedical Engineering*, vol. 47, no. 9, pp. 1185–1194, 2000.

[182] A. Ismail-Fawaz, M. Devanne, S. Berretti, J. Weber, and G. Forestier, "Look into the LITE in deep learning for time series classification", en, *International Journal of Data Science and Analytics*, vol. 20, no. 4, pp. 4029–4049, Oct. 2025, ISSN: 2364-4168. DOI: 10.1007/s41060-024-00708-5. [Online]. Available: https://doi.org/10.1007/s41060-024-00708-5.

[183] N. M. Foumani, C. W. Tan, G. I. Webb, and M. Salehi, "Improving position encoding of transformers for multivariate time series classification", en, *Data Mining and Knowledge Discovery*, vol. 38, no. 1, pp. 22–48, Jan. 2024, ISSN: 1573-756X. DOI: 10.1007/s10618-023-00948-2. [Online]. Available: https://doi.org/10.1007/s10618-023-00948-2.

[184] M.-J. Joe and S.-C. Pyo, "Classification of Sleep Stage with Biosignal Images Using Convolutional Neural Networks", en, *Applied Sciences*, vol. 12, no. 6, Mar. 2022, ISSN: 2076-3417. DOI: 10.3390/app12063028. [Online]. Available: https://www.mdpi.com/2076-3417/12/6/3028.

[185] W. Huang, Y. Wang, H. Cheng, *et al.*, "A unified time-frequency foundation model for sleep decoding", en, *Nature Communications*, vol. 17, no. 1, p. 1198, Jan. 2026, ISSN: 2041-1723. DOI: 10.1038/s41467-025-67970-4. [Online]. Available: https://www.nature.com/articles/s41467-025-67970-4.

[186] Y. Guo, M. Nowakowski, and W. Dai, "FlexSleepTransformer: A transformer-based sleep staging model with flexible input channel configurations", *Scientific Reports*, vol. 14, p. 26 312, Nov. 2024, ISSN: 2045-2322. DOI: 10.1038/s41598-024-76197-0. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC11530688/.

[187] P. Somer, "Algorithmic accountability of low-code/no-code artificial intelligence: A literature review", *Thirty-first Americas Conference on Information Systems, Montréal, 2025*, 2025.

[188] H. Herath, H. Herath, B. Madhusanka, and L. Guruge, "Data protection challenges in the processing of sensitive data", in *Data Protection: The Wake of AI and Machine Learning*, Springer, 2024, pp. 155–179.

[189] A. Cavoukian *et al.*, "Privacy by design: The 7 foundational principles", *Information and privacy commissioner of Ontario, Canada*, vol. 5, no. 2009, p. 12, 2009.

[190] A. Tornede, D. Deng, T. Eimer, *et al.*, "Automl in the age of large language models: Current challenges, future opportunities and risks", *Transactions on Machine Learning Research*, 2023.

[191] J. Amann, A. Blasimme, E. Vayena, D. Frey, V. I. Madai, and the Precise4Q consortium, "Explainability for artificial intelligence in healthcare: A multidisciplinary perspective", *BMC Medical Informatics and Decision Making*, vol. 20, no. 1, p. 310, Nov. 2020, ISSN: 1472-6947. DOI: 10.1186/s12911-020-01332-6. [Online]. Available: https://doi.org/10.1186/s12911-020-01332-6.

[192] S. Labkoff, B. Oladimeji, J. Kannry, *et al.*, "Toward a responsible future: Recommendations for AI-enabled clinical decision support", en, *Journal of the American Medical Informatics Association*, vol. 31, no. 11, pp. 2730–2739, Nov. 2024, ISSN: 1067-5027, 1527-974X. DOI: 10.1093/jamia/ocae209. [Online]. Available: https://academic.oup.com/jamia/article/31/11/2730/7776823.

[193] R. S. Rosenberg and S. Van Hout, "The American Academy of Sleep Medicine inter-scorer reliability program: Sleep stage scoring", eng, *Journal of clinical sleep medicine: JCSM: official publication of the American Academy of Sleep Medicine*, vol. 9, no. 1, pp. 81–87, Jan. 2013, ISSN: 1550-9397. DOI: 10.5664/jcsm.2350.

[194] Y. J. Lee, J. Y. Lee, J. H. Cho, and J. H. Choi, "Interrater reliability of sleep stage scoring: A meta-analysis", *Journal of Clinical Sleep Medicine : JCSM : Official Publication of the American Academy of Sleep Medicine*, vol. 18, no. 1, pp. 193–202, Jan. 2022, ISSN: 1550-9389. DOI: 10.5664/jcsm.9538. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC8807917/.

[195] J. P. Bakker, M. Ross, A. Cerny, *et al.*, "Scoring sleep with artificial intelligence enables quantification of sleep stage ambiguity: Hypnodensity based on multiple expert scorers and auto-scoring", *Sleep*, vol. 46, no. 2, zsac154, Feb. 2023, ISSN: 0161-8105. DOI: 10.1093/sleep/zsac154. [Online]. Available: https://doi.org/10.1093/sleep/zsac154.

[196] 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", English, *Frontiers in Sleep*, vol. 2, Apr. 2023, ISSN: 2813-2890. DOI: 10.3389/frsle.2023.1163477. [Online]. Available: https://www.frontiersin.org/journals/sleep/articles/10.3389/frsle.2023.1163477/full.

[197] M. Braun, M. Stockhoff, M. Tijssen, S. Dietz-Terjung, S. Coughlin, and C. Schöbel, "A Systematic Review on the Technical Feasibility of Home-Polysomnography for Diagnosis of Sleep Disorders in Adults", en, *Current Sleep Medicine Reports*, vol. 10, no. 2, pp. 276–288, Jun. 2024, ISSN: 2198-6401. DOI: 10.1007/s40675-024-00301-z. [Online]. Available: https://doi.org/10.1007/s40675-024-00301-z.

[198] M. Alattar, A. Govind, and S. Mainali, "Artificial Intelligence Models for the Automation of Standard Diagnostics in Sleep Medicine—A Systematic Review", en, *Bioengineering*, vol. 11, no. 3, p. 206, Mar. 2024, Number: 3, ISSN: 2306-5354. DOI: 10.3390/bioengineering11030206. [Online]. Available: https://www.mdpi.com/2306-5354/11/3/206.

[199] 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, pp. 247–272, Jul. 2023, ISSN: 2093-9868. DOI: 10.1007/s13534-023-00299-3. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10382458/.

[200] M. Perslev, S. Darkner, L. Kempfner, M. Nikolic, P. J. Jennum, and C. Igel, "U-Sleep: Resilient high-frequency sleep staging", en, *npj Digital Medicine*, vol. 4, no. 1, p. 72, Apr. 2021, ISSN: 2398-6352. DOI: 10.1038/s41746-021-00440-5. [Online]. Available: https://www.nature.com/articles/s41746-021-00440-5.

[201] R. Vallat and M. P. Walker, "An open-source, high-performance tool for automated sleep staging", *eLife*, vol. 10, A. Peyrache, C. Büchel, and S. Bagur, Eds., e70092, Oct. 2021, ISSN: 2050-084X. DOI: 10.7554/eLife.70092. [Online]. Available: https://doi.org/10.7554/eLife.70092.

[202] H. Phan, K. P. Lorenzen, E. Heremans, *et al.*, "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, Oct. 2023, ISSN: 2168-2208. DOI: 10.1109/JBHI.2023.3303197. [Online]. Available: https://ieeexplore.ieee.org/document/10210638.

[203] M. Yazdi, M. Samaee, and D. Massicotte, "A Review on Automated Sleep Study", en, *Annals of Biomedical Engineering*, vol. 52, no. 6, pp. 1463–1491, Jun. 2024, ISSN: 1573-9686. DOI: 10.1007/s10439-024-03486-0. [Online]. Available: https://doi.org/10.1007/s10439-024-03486-0.

[204] E. Larson, A. Gramfort, D. A. Engemann, *et al.*, *MNE-Python*, Nov. 2025. DOI: 10.5281/ZENODO.592483. [Online]. Available: https://zenodo.org/doi/10.5281/zenodo.592483.

[205] P. Welch, "The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms", *IEEE Transactions on Audio and Electroacoustics*, vol. 15, no. 2, pp. 70–73, Jun. 1967, ISSN: 1558-2582. DOI: 10.1109/TAU.1967.1161901. [Online]. Available: https://ieeexplore.ieee.org/document/1161901.

[206] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, *The Curious Case of Neural Text Degeneration*, arXiv:1904.09751 [cs], Feb. 2020. DOI: 10.48550/arXiv.1904.09751. [Online]. Available: http://arxiv.org/abs/1904.09751.

[207] E. S. Arnardottir, A. S. Islind, and M. Óskarsdóttir, "The Future of Sleep Measurements: A Review and Perspective", eng, *Sleep Medicine Clinics*, vol. 16, no. 3, pp. 447–464, Sep. 2021, ISSN: 1556-4088. DOI: 10.1016/j.jsmc.2021.05.004.

[208] M. de Zambotti, N. Cellini, A. Goldstone, I. M. Colrain, and F. C. Baker, "Wearable Sleep Technology in Clinical and Research Settings", eng, *Medicine and Science in Sports and Exercise*, vol. 51, no. 7, pp. 1538–1557, Jul. 2019, ISSN: 1530-0315. DOI: 10.1249/MSS.0000000000001947.

[209] R. B. Berry, S. F. Quan, A. R. Abreu, *et al.*, *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications.* 2020.

[210] H. van Gorp, I. A. M. Huijben, P. Fonseca, R. J. G. van Sloun, S. Overeem, and M. M. van Gilst, "Certainty about uncertainty in sleep staging: A theoretical framework", *Sleep*, vol. 45, no. 8, zsac134, Aug. 2022, ISSN: 0161-8105. DOI: 10.1093/sleep/zsac134. [Online]. Available: https://doi.org/10.1093/sleep/zsac134.

[211] 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", en, *npj Digital Medicine*, vol. 4, no. 1, pp. 1–9, Sep. 2021, Number: 1, ISSN: 2398-6352. DOI: 10.1038/s41746-

021-00515-3. [Online]. Available: https://www.nature.com/articles/s41746-021-00515-3.

[212] G. M. Anderson, "Error propagation by the Monte Carlo method in geochemical calculations", en, *Geochimica et Cosmochimica Acta*, vol. 40, no. 12, pp. 1533–1538, Dec. 1976, ISSN: 0016-7037. DOI: 10.1016/0016-7037(76)90092-2. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0016703776900922.

[213] J. Kager and C. Herwig, "Monte Carlo-Based Error Propagation for a More Reliable Regression Analysis across Specific Rates in Bioprocesses", *Bioengineering*, vol. 8, no. 11, p. 160, Oct. 2021, ISSN: 2306-5354. DOI: 10.3390/bioengineering8110160. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8614739/.

[214] Y. Li, S. Zhang, R. Hobbs, *et al.*, "Monte Carlo sampling for error propagation in linear regression and applications in isochron geochronology", *Science Bulletin*, vol. 64, Feb. 2019. DOI: 10.1016/j.scib.2018.12.019.

[215] M. Óskarsdóttir, A. S. Islind, E. August, E. S. Arnardóttir, F. Patou, and A. M. Maier, "Importance of Getting Enough Sleep and Daily Activity Data to Assess Variability: Longitudinal Observational Study", EN, *JMIR Formative Research*, vol. 6, no. 2, e31807, Feb. 2022, Company: JMIR Formative Research Distributor: JMIR Formative Research Institution: JMIR Formative Research Label: JMIR Formative Research. DOI: 10.2196/31807. [Online]. Available: https://formative.jmir.org/2022/2/e31807.

[216] *Wearable device market share by category worldwide 2021*, en. [Online]. Available: https://www.statista.com/statistics/690731/wearables-worldwide-shipments-by-product-category/.

[217] *Global smartwatch shipments market share 2022*, en. [Online]. Available: https://www.statista.com/statistics/910862/worldwide-smartwatch-shipment-market-share/.

[218] S. Khosla, M. C. Deak, D. Gault, *et al.*, "Consumer Sleep Technology: An American Academy of Sleep Medicine Position Statement", *Journal of Clinical Sleep Medicine*, vol. 14, no. 05, pp. 877–880, 2018. DOI: 10.5664/jcsm.7128. [Online]. Available: https://jcsm.aasm.org/doi/10.5664/jcsm.7128.

[219] C. M. Depner, P. C. Cheng, J. K. Devine, *et al.*, "Wearable technologies for developing sleep and circadian biomarkers: A summary of workshop discussions", *Sleep*, vol. 43, no. 2, zsz254, Feb. 2020, ISSN: 0161-8105. DOI: 10.1093/sleep/zsz254. [Online]. Available: https://doi.org/10.1093/sleep/zsz254.

[220] L. Menghini, N. Cellini, A. Goldstone, F. C. Baker, and M. de Zambotti, "A standardized framework for testing the performance of sleep-tracking technology: Step-by-step guidelines and open-source code", *Sleep*, vol. 44, no. 2, zsaa170, Feb. 2021, ISSN: 0161-8105. DOI: 10.1093/sleep/zsaa170. [Online]. Available: https://doi.org/10.1093/sleep/zsaa170.

[221] L. E. Rentz, H. K. Ulman, and S. M. Galster, "Deconstructing Commercial Wearable Technology: Contributions toward Accurate and Free-Living Monitoring of Sleep", en, *Sensors*, vol. 21, no. 15, p. 5071, Jan. 2021, Number: 15, ISSN: 1424-8220. DOI: 10.3390/s21155071. [Online]. Available: https://www.mdpi.com/1424-8220/21/15/5071.

[222] M. de Zambotti, N. Cellini, L. Menghini, M. Sarlo, and F. C. Baker, "Sensors Capabilities, Performance, and Use of Consumer Sleep Technology", eng, *Sleep Medicine Clinics*, vol. 15, no. 1, pp. 1–30, Mar. 2020, ISSN: 1556-4088. DOI: 10.1016/j.jsmc.2019.11.003.

[223] S. A. Imtiaz, "A Systematic Review of Sensing Technologies for Wearable Sleep Staging", en, *Sensors*, vol. 21, no. 5, p. 1562, Jan. 2021, Number: 5, ISSN: 1424-8220. DOI: 10.3390/s21051562. [Online]. Available: https://www.mdpi.com/1424-8220/21/5/1562.

[224] M. R. Lujan, I. Perez-Pozuelo, and M. A. Grandner, "Past, Present, and Future of Multisensory Wearable Technology to Monitor Sleep and Circadian Rhythms", *Frontiers in Digital Health*, vol. 3, 2021, ISSN: 2673-253X. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fdgth.2021.721919.

[225] Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T. H. Falk, and J. Faubert, "Deep learning-based electroencephalography analysis: A systematic review", en, *Journal of Neural Engineering*, vol. 16, no. 5, p. 051 001, Aug. 2019, ISSN: 1741-2552. DOI: 10.1088/1741-2552/ab260c. [Online]. Available: https://doi.org/10.1088/1741-2552/ab260c.

[226] S. Gashi, L. Alecci, E. D. Lascio, M. E. Debus, F. Gasparini, and S. Santini, "The Role of Model Personalization for Sleep Stage and Sleep Quality Recognition Using Wearables", *IEEE Pervasive Computing*, vol. 21, no. 2, pp. 69–77, Apr. 2022, Conference Name: IEEE Pervasive Computing, ISSN: 1558-2590. DOI: 10.1109/MPRV.2022.3164334.

[227] B. M. Wulterkens, P. Fonseca, L. W. Hermans, *et al.*, "It is All in the Wrist: Wearable Sleep Staging in a Clinical Population versus Reference Polysomnography", English, *Nature and Science of Sleep*, vol. 13, pp. 885–897, Jun. 2021. DOI: 10.2147/NSS.S306808. [Online]. Available: https://www.dovepress.com/it-is-all-in-the-wrist-wearable-sleep-staging-in-a-clinical-population-peer-reviewed-fulltext-article-NSS.

[228] P. Fonseca, T. Weysen, M. S. Goelema, *et al.*, "Validation of Photoplethysmography-Based Sleep Staging Compared With Polysomnography in Healthy Middle-Aged Adults", *Sleep*, vol. 40, no. 7, zsx097, Jul. 2017, ISSN: 0161-8105. DOI: 10.1093/sleep/zsx097. [Online]. Available: https://doi.org/10.1093/sleep/zsx097.

[229] A. S. Anusha, S. P. Preejith, T. J. Akl, and M. Sivaprakasam, "Electrodermal activity based autonomic sleep staging using wrist wearable", en, *Biomedical Signal Processing and Control*, vol. 75, p. 103 562, May 2022, ISSN: 1746-8094. DOI: 10.1016/j.bspc.2022.103562. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809422000842.

[230] Y.-F. Ko, P.-H. Kuo, C.-F. Wang, *et al.*, "Quantification Analysis of Sleep Based on Smartwatch Sensors for Parkinson's Disease", en, *Biosensors*, vol. 12, no. 2, p. 74, Feb. 2022, Number: 2, ISSN: 2079-6374. DOI: 10.3390/bios12020074. [Online]. Available: https://www.mdpi.com/2079-6374/12/2/74.

[231] M. Radha, P. Fonseca, A. Moreau, *et al.*, "A deep transfer learning approach for wearable sleep stage classification with photoplethysmography", en, *npj Digital Medicine*, vol. 4, no. 1, p. 135, Dec. 2021, ISSN: 2398-6352. DOI: 10.1038/s41746-021-00510-8. [Online]. Available: https://www.nature.com/articles/s41746-021-00510-8.

[232] Q. Li, Q. Li, A. S. Cakmak, *et al.*, "Transfer learning from ECG to PPG for improved sleep staging from wrist-worn wearables", en, *Physiological Measurement*, vol. 42, no. 4, p. 044 004, May 2021, ISSN: 0967-3334. DOI: 10.1088/1361-6579/abf1b0. [Online]. Available: https://dx.doi.org/10.1088/1361-6579/abf1b0.

[233] K. Sundararajan, S. Georgievska, B. H. W. te Lindert, *et al.*, "Sleep classification from wrist-worn accelerometer data using random forests", en, *Scientific Reports*, vol. 11, no. 1, p. 24, Jan. 2021, Number: 1, ISSN: 2045-2322. DOI: 10.1038/s41598-020-79217-x. [Online]. Available: https://www.nature.com/articles/s41598-020-79217-x.

[234] O. Walch, Y. Huang, D. Forger, and C. Goldstein, "Sleep stage prediction with raw acceleration and photoplethysmography heart rate data derived from a consumer wearable device", *Sleep*, vol. 42, no. 12, zsz180, Dec. 2019, ISSN: 0161-8105. DOI: 10.1093/sleep/zsz180. [Online]. Available: https://doi.org/10.1093/sleep/zsz180.

[235] G. Brupbacher, D. Straus, H. Porschke, *et al.*, "The acute effects of aerobic exercise on sleep in patients with depression: Study protocol for a randomized controlled trial", *Trials*, vol. 20, no. 1, p. 352, Jun. 2019, ISSN: 1745-6215. DOI: 10.1186/s13063-019-3415-3. [Online]. Available: https://doi.org/10.1186/s13063-019-3415-3.

[236] J. Zhang, "Modern Monte Carlo methods for efficient uncertainty quantification and propagation: A survey", en, *WIREs Computational Statistics*, vol. 13, no. 5, e1539, 2021, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wics.1539, ISSN: 1939-0068. DOI: 10.1002/wics.1539. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/wics.1539.

[237] P. R. G. Couto, J. C. Damasceno, S. P. d. Oliveira, P. R. G. Couto, J. C. Damasceno, and S. P. d. Oliveira, *Monte Carlo Simulations Applied to Uncertainty in Measurement*, en. IntechOpen, Mar. 2013, Publication Title: Theory and Applications of Monte Carlo Simulations, ISBN: 978-953-51-1012-5. DOI: 10.5772/53014. [Online]. Available: https://www.intechopen.com/state.item.id.

[238] T. P. Morris, I. R. White, and M. J. Crowther, "Using simulation studies to evaluate statistical methods", eng, *Statistics in Medicine*, vol. 38, no. 11, pp. 2074–2102, May 2019, ISSN: 1097-0258. DOI: 10.1002/sim.8086.

[239] B. Kemp, A. H. Zwinderman, B. Tuk, H. A. Kamphuisen, and J. J. Oberye, "Analysis of a sleep-dependent neuronal feedback loop: The slow-wave microcontinuity of the EEG", *IEEE Transactions on Biomedical Engineering*, vol. 47, no. 9, pp. 1185–1194, 2000.

[240] A. L. Goldberger, L. A. Amaral, L. Glass, *et al.*, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals", eng, *Circulation*, vol. 101, no. 23, E215–220, Jun. 2000, ISSN: 1524-4539. DOI: 10.1161/01.cir.101.23.e215.

[241] Q. Li, Q. Li, C. Liu, S. P. Shashikumar, S. Nemati, and G. D. Clifford, "Deep learning in the cross-time frequency domain for sleep staging from a single-lead electrocardiogram", en, *Physiological Measurement*, vol. 39, no. 12, p. 124 005, Dec. 2018, ISSN: 0967-3334. DOI: 10.1088/1361-6579/aaf339. [Online]. Available: https://doi.org/10.1088/1361-6579/aaf339.

[242] K. Krippendorff, *Content analysis: an introduction to its methodology*, eng, 2. ed., [Nachdr.] Thousand Oaks, Calif.: Sage Publ, 2004, ISBN: 978-0-7619-1545-4 978-0-7619-1544-7.

[243] J. L. Fleiss, "Measuring nominal scale agreement among many raters", *Psychological Bulletin*, vol. 76, no. 5, pp. 378–382, 1971, ISSN: 1939-1455. DOI: 10.1037/h0031619.

[244] P. E. Shrout and J. L. Fleiss, "Intraclass correlations: Uses in assessing rater reliability", *Psychological Bulletin*, vol. 86, no. 2, pp. 420–428, 1979, ISSN: 1939-1455. DOI: 10.1037/0033-2909.86.2.420.

[245] K. O. McGraw and S. P. Wong, "Forming inferences about some intraclass correlation coefficients", *Psychological Methods*, vol. 1, no. 1, pp. 30–46, 1996, ISSN: 1939-1463. DOI: 10.1037/1082-989X.1.1.30.

[246] T. K. Koo and M. Y. Li, "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research", *Journal of Chiropractic Medicine*, vol. 15, no. 2, pp. 155–163, Jun. 2016, ISSN: 1556-3707. DOI: 10.1016/j.jcm.2016.02.012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1556370716000158.

[247] M. I. Boulos, T. Jairam, T. Kendzerska, J. Im, A. Mekhael, and B. J. Murray, "Normal polysomnography parameters in healthy adults: A systematic review and meta-analysis", *The Lancet Respiratory Medicine*, vol. 7, no. 6, pp. 533–543, Jun. 2019, ISSN: 2213-2600. DOI: 10.1016/S2213-2600(19)30057-8. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2213260019300578.

[248] F. Cabitza, A. Locoro, C. Alderighi, R. Rasoini, D. Compagnone, and P. Berjano, "The elephant in the record: On the multiplicity of data recording work", EN, *Health Informatics Journal*, vol. 25, no. 3, pp. 475–490, Sep. 2019, ISSN: 1460-4582. DOI: 10.1177/1460458218824705. [Online]. Available: https://doi.org/10.1177/1460458218824705.

[249] E. Tasali, R. Leproult, D. A. Ehrmann, and E. Van Cauter, "Slow-wave sleep and the risk of type 2 diabetes in humans", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 3, pp. 1044–1049, Jan. 2008, ISSN: 0027-8424. DOI: 10.1073/pnas.0706446105. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC2242689/.

[250] M. M. Fung, K. Peters, S. Redline, *et al.*, "DECREASED SLOW WAVE SLEEP INCREASES RISK OF DEVELOPING HYPERTENSION IN ELDERLY MEN", *Hypertension*, vol. 58, no. 4, pp. 596–603, Oct. 2011, ISSN: 0194-911X. DOI: 10.1161/HYPERTENSIONAHA.111.174409. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC3176739/.

[251] S. Javaheri, Y. Y. Zhao, N. M. Punjabi, S. F. Quan, D. J. Gottlieb, and S. Redline, "Slow-Wave Sleep Is Associated With Incident Hypertension: The Sleep Heart Health Study", *Sleep*, vol. 41, no. 1, zsx179, Oct. 2017, ISSN: 0161-8105. DOI: 10.1093/sleep/zsx179. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC5806562/.

[252] B. A. Mander, V. Rao, B. Lu, *et al.*, "Prefrontal atrophy, disrupted NREM slow waves, and impaired hippocampal-dependent memory in aging", *Nature neuroscience*, vol. 16, no. 3, pp. 357–364, Mar. 2013, ISSN: 1097-6256. DOI: 10.1038/nn.3324. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC4286370/.

[253] H. Jahrami, W. Husain, K. Trabelsi, *et al.*, "Artificial intelligence and sleep medicine II: A scoping review of applications, advancements, and future directions", *Sleep Medicine Reviews*, vol. 85, p. 102 212, Feb. 2026, ISSN: 1087-0792. DOI: 10.1016/j.smrv.2025.102212. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1087079225001650.

[254] N. F. Watson and C. R. Fernandez, "Artificial intelligence and sleep: Advancing sleep medicine", *Sleep Medicine Reviews*, vol. 59, p. 101 512, Oct. 2021, ISSN: 1087-0792. DOI: 10.1016/j.smrv.2021.101512. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1087079221000976.

[255] J. B. Stephansen, A. N. Olesen, M. Olsen, *et al.*, "Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy", en, *Nature Communications*, vol. 9, no. 1, p. 5229, Dec. 2018, ISSN: 2041-1723. DOI: 10.1038/s41467-018-07229-3. [Online]. Available: https://www.nature.com/articles/s41467-018-07229-3.

[256] V. T. v. Hees, S. Sabia, K. N. Anderson, *et al.*, "A Novel, Open Access Method to Assess Sleep Duration Using a Wrist-Worn Accelerometer", en, *PLOS ONE*, vol. 10, no. 11, e0142533, Nov. 2015, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0142533. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0142533.

[257] V. T. van Hees, S. Sabia, S. E. Jones, *et al.*, "Estimating sleep parameters using an accelerometer without sleep diary", en, *Scientific Reports*, vol. 8, no. 1, p. 12 975, Aug. 2018, ISSN: 2045-2322. DOI: 10.1038/s41598-018-31266-z. [Online]. Available: https://www.nature.com/articles/s41598-018-31266-z.

[258] D. C. Lim, D. R. Mazzotti, K. Sutherland, *et al.*, "Reinventing polysomnography in the age of precision medicine", *Sleep Medicine Reviews*, vol. 52, p. 101 313, Aug. 2020, ISSN: 1087-0792. DOI: 10.1016/j.smrv.2020.101313. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1087079220300563.

[259] B. Lechat, H. Scott, G. Naik, *et al.*, "New and Emerging Approaches to Better Define Sleep Disruption and Its Consequences", English, *Frontiers in Neuroscience*, vol. 15, Oct. 2021, ISSN: 1662-453X. DOI: 10.3389/fnins.2021.751730. [Online]. Available: https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2021.751730/full.

[260] J. Guo and Y. Xiao, "New Metrics from Polysomnography: Precision Medicine for OSA Interventions", *Nature and Science of Sleep*, vol. 15, pp. 69–77, Dec. 2023, _eprint: https://www.tandfonline.com/doi/pdf/10.2147/NSS.S400048, ISSN: null. DOI: 10.2147/NSS.S400048. [Online]. Available: https://www.tandfonline.com/doi/abs/10.2147/NSS.S400048.

[261] Y. Liu, A. A. Ghafoor, M. Hajipour, and N. Ayas, "Role of precision medicine in obstructive sleep apnoea", en, *BMJ Medicine*, vol. 2, no. 1, Jan. 2023, ISSN: 2754-0413. DOI: 10.1136/bmjmed-2022-000218. [Online]. Available: https://bmjmedicine.bmj.com/content/2/1/e000218.

[262] M. Younes, M. Ostrowski, M. Soiferman, *et al.*, "Odds Ratio Product of Sleep EEG as a Continuous Measure of Sleep State", *Sleep*, vol. 38, no. 4, pp. 641–654, Apr. 2015, ISSN: 0161-8105. DOI: 10.5665/sleep.4588. [Online]. Available: https://doi.org/10.5665/sleep.4588.

[263] M. Younes, "The case for using digital EEG analysis in clinical sleep medicine", en, *Sleep Science and Practice*, vol. 1, no. 1, p. 2, Feb. 2017, ISSN: 2398-2683. DOI: 10.1186/s41606-016-0005-0. [Online]. Available: https://doi.org/10.1186/s41606-016-0005-0.

[264] M. Younes, B. Gerardy, A. I. Pack, S. T. Kuna, C. Castro-Diehl, and S. Redline, "Sleep architecture based on sleep depth and propensity: Patterns in different demographics and sleep disorders and association with health outcomes", en, *Sleep*, vol. 45, no. 6, Jun. 2022, ISSN: 0161-8105. DOI: 10.1093/sleep/zsac059. [Online]. Available: https://dx.doi.org/10.1093/sleep/zsac059.

[265] Y. Leng, J. Chen, C. Cavaillès, *et al.*, "Sleep Temporal Entropy as a Novel Digital Biomarker of Sleep Fragmentation for Cardiometabolic and Mortality Risk", *Research Square*, rs.3.rs–7 433 027, Sep. 2025, ISSN: 2693-5015. DOI: 10.21203/rs.3.rs-7433027/v1. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC12425052/.

[266] *Methods for Obtaining and Analyzing Unattended Polysomnography Data for a Multicenter Study | SLEEP | Oxford Academic*. [Online]. Available: https://academic.oup.com/sleep/article/21/7/759/2731681.

[267] F. Chung, H. R. Abdullah, and P. Liao, "STOP-Bang Questionnaire: A Practical Approach to Screen for Obstructive Sleep Apnea", *Chest*, vol. 149, no. 3, pp. 631–638, Mar. 2016, ISSN: 0012-3692. DOI: 10.1378/chest.15-0903. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0012369215000185.

[268] C. M. Morin, G. Belleville, L. Bélanger, and H. Ivers, "The Insomnia Severity Index: Psychometric Indicators to Detect Insomnia Cases and Evaluate Treatment Response", en, *Sleep*, vol. 34, no. 5, pp. 601–608, May 2011, ISSN: 0161-8105. DOI: 10.1093/sleep/34.5.601. [Online]. Available: https://dx.doi.org/10.1093/sleep/34.5.601.

[269] "Validation of the International Restless Legs Syndrome Study Group rating scale for restless legs syndrome", *Sleep Medicine*, Restless Legs Syndrome, vol. 4, no. 2, pp. 121–132, Mar. 2003, ISSN: 1389-9457. DOI: 10.1016/S1389-9457(02)00258-7. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1389945702002587.

[270] J. Van Der Donckt, J. Van Der Donckt, E. Deprost, *et al.*, *Do Not Sleep on Linear Models: Simple and Interpretable Techniques Outperform Deep Learning for Sleep Scoring*, arXiv:2207.07753 [cs, eess, stat], Jul. 2022. DOI: 10.48550/arXiv.2207.07753. [Online]. Available: http://arxiv.org/abs/2207.07753.

[271] A. Supratak and Y. Guo, "TinySleepNet: An Efficient Deep Learning Model for Sleep Stage Scoring based on Raw Single-Channel EEG", in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, Jul. 2020, pp. 641–644. DOI: 10.1109/EMBC44109.2020.9176741.

[272] Y. Hong, "On computing the distribution function for the Poisson binomial distribution", *Computational Statistics & Data Analysis*, vol. 59, pp. 41–51, Mar. 2013, ISSN: 0167-9473. DOI: 10.1016/j.csda.2012.10.006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167947312003568.

[273] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*, arXiv:2010.11929 [cs], Jun. 2021. DOI: 10.48550/arXiv.2010.11929. [Online]. Available: http://arxiv.org/abs/2010.11929.

[274] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked Autoencoders Are Scalable Vision Learners", en, 2022, pp. 16 000–16 009. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2022/html/He_Masked_Autoencoders_Are_Scalable_Vision_Learners_CVPR_2022_paper.

[275] R. D. Riley, L. Archer, K. I. E. Snell, *et al.*, "Evaluation of clinical prediction models (part 2): How to undertake an external validation study", *The BMJ*, vol. 384, e074820, Jan. 2024, ISSN: 0959-8138. DOI: 10.1136/bmj-2023-074820. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC10788734/.

[276] F. Cabitza, A. Campagner, D. Albano, *et al.*, "The Elephant in the Machine: Proposing a New Metric of Data Reliability and its Application to a Medical Case to Assess Classification Reliability", en, *Applied Sciences*, vol. 10, no. 11, p. 4014, Jan. 2020, ISSN: 2076-3417. DOI: 10.3390/app10114014. [Online]. Available: https://www.mdpi.com/2076-3417/10/11/4014.

[277] R. Thapa, M. R. Kjaer, B. He, *et al.*, "A multimodal sleep foundation model for disease prediction", en, *Nature Medicine*, vol. 32, no. 2, pp. 752–762, Feb. 2026, ISSN: 1546-170X. DOI: 10.1038/s41591-025-04133-4. [Online]. Available: https://www.nature.com/articles/s41591-025-04133-4.

[278] A. Shollo, K. Hopf, T. Thiess, and O. Müller, "Shifting ml value creation mechanisms: A process model of ml value creation", *The Journal of Strategic Information Systems*, vol. 31, no. 3, p. 101 734, 2022.

[279] S. Faraj, S. Pachidi, and K. Sayegh, "Working and organizing in the age of the learning algorithm", *Information and Organization*, vol. 28, no. 1, pp. 62–70, 2018.

[280] T. Grønsund and M. Aanestad, "Augmenting the algorithm: Emerging human-in-the-loop work configurations", *The Journal of Strategic Information Systems*, vol. 29, no. 2, p. 101 614, 2020.

[281] A. R. Dennis, A. Lakhiwal, and A. Sachdeva, "Ai agents as team members: Effects on satisfaction, conflict, trustworthiness, and willingness to work with", *Journal of Management Information Systems*, vol. 40, no. 2, pp. 307–337, 2023.

[282] A. Fügener, J. Grahl, A. Gupta, and W. Ketter, "Cognitive challenges in human–artificial intelligence collaboration: Investigating the path toward productive delegation", *Information Systems Research*, vol. 33, no. 2, pp. 678–696, 2022.

[283] M. Vössing, N. Kühl, M. Lind, and G. Satzger, "Designing transparency for effective human-ai collaboration", *Information Systems Frontiers*, vol. 24, no. 3, pp. 877–895, 2022.

[284] H. Vallo Hult, A. S. Islind, and L. Norström, "Reconfiguring professionalism in digital work", *Systems, Signs & Actions: An International Journal on Information Technology, Action, Communication and Workpractices*, vol. 12, pp. 1–17, 2021.

[285] N. Berente, B. Gu, J. Recker, and R. Santhanam, "Managing artificial intelligence.", *MIS quarterly*, vol. 45, no. 3, 2021.

[286] S. Ali, T. Abuhmed, S. El-Sappagh, *et al.*, "Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence", *Information Fusion*, vol. 99, p. 101 805, Nov. 2023, ISSN: 1566-2535. DOI: 10.1016/j.inffus.2023.101805. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253523001148.

[287] D. Castelvecchi, "Can we open the black box of ai?", *Nature News*, vol. 538, no. 7623, p. 20, 2016.

[288] L. Hylving, D. Koutsikouri, J. Bornemark, and S. Lindberg, "Ratio and intellectus: Towards a conceptual framework for understanding human and artificial intelligence", in *International Conference on Information Systems (ICIS), ICIS 2022, Copenhagen, Denmark, 9-14 December, 2022*, Association for Information Systems, 2022.

[289] S. Nyholm, "Artificial intelligence and human enhancement: Can ai technologies make us more (artificially) intelligent?", *Cambridge Quarterly of Healthcare Ethics*, vol. 33, no. 1, pp. 76–88, 2024.

[290] B. Kovalerchuk, M. A. Ahmad, and A. Teredesai, "Survey of explainable machine learning with visual and granular methods beyond quasi-explanations", *Interpretable artificial intelligence: A perspective of granular computing*, pp. 217–267, 2021.

[291] J. P. Cornelissen and M. D. Werner, "Putting framing in perspective: A review of framing and frame analysis across the management and organizational literature", *Academy of Management Annals*, vol. 8, no. 1, pp. 181–235, 2014.

[292] B. Gray, J. M. Purdy, and S. Ansari, "From interactions to institutions: Microprocesses of framing and mechanisms for the structuring of institutional fields", *Academy of management review*, vol. 40, no. 1, pp. 115–143, 2015.

[293] F. Ivarsson, "Applying Framing Theory in Digital Transformation Research: Suggestions for Future Research", in *Hawaii International Conference on System Sciences 2022 (HICSS-55)*, Jan. 2022. [Online]. Available: https://aisel.aisnet.org/hicss-55/os/innovation/3.

[294] D. Vale, A. El-Sharif, and M. Ali, "Explainable artificial intelligence (XAI) post-hoc explainability methods: Risks and limitations in non-discrimination law", en, *AI and Ethics*, vol. 2, no. 4, pp. 815–826, Nov. 2022, ISSN: 2730-5961. DOI: 10.1007/s43681-022-00142-y. [Online]. Available: https://doi.org/10.1007/s43681-022-00142-y.

[295] W. Jin, X. Li, M. Fatehi, and G. Hamarneh, "Guidelines and evaluation of clinical explainable AI in medical image analysis", *Medical Image Analysis*, vol. 84, p. 102 684, Feb. 2023, ISSN: 1361-8415. DOI: 10.1016/j.media.2022.102684. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841522003127.

[296] A. Saranya and R. Subhashini, "A systematic review of Explainable Artificial Intelligence models and applications: Recent developments and future trends", *Decision Analytics Journal*, vol. 7, p. 100 230, Jun. 2023, ISSN: 2772-6622. DOI: 10.1016/j.dajour.2023.100230. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S277266222300070X.

[297] M. Merry, P. Riddle, and J. Warren, "A mental models approach for defining explainable artificial intelligence", *BMC Medical Informatics and Decision Making*, vol. 21, no. 1, p. 344, Dec. 2021, ISSN: 1472-6947. DOI: 10.1186/s12911-021-01703-7. [Online]. Available: https://doi.org/10.1186/s12911-021-01703-7.

[298] E. Goffman, *Frame analysis: An essay on the organization of experience.* Harvard University Press, 1974.

[299] A. Essén and S. W. Värlander, "How materiality enables and constrains framing practices: Affordances of a rheumatology e-service", *Journal of Management Inquiry*, vol. 28, no. 4, pp. 458–471, 2019.

[300] W. J. Orlikowski and D. C. Gash, "Technological frames: Making sense of information technology in organizations", *ACM Transactions on Information Systems (TOIS)*, vol. 12, no. 2, pp. 174–207, 1994.

[301] J. Purdy, S. Ansari, and B. Gray, "Are logics enough? framing as an alternative tool for understanding institutional meaning making", *Journal of Management Inquiry*, vol. 28, no. 4, pp. 409–419, 2019.

[302] E. Davidson, "A technological frames perspective on information technology and organizational change", *The journal of applied behavioral science*, vol. 42, no. 1, pp. 23–39, 2006.

[303] J. Hoppmann, L. D. Anadon, and V. Narayanamurti, "Why matter matters: How technology characteristics shape the strategic framing of technologies", *Research Policy*, vol. 49, no. 1, p. 103 882, 2020.

[304] B. W. Young, L. Mathiassen, and E. Davidson, "Inconsistent and incongruent frames during it-enabled change: An action research study into sales process innovation", *Journal of the Association for Information Systems*, vol. 17, no. 7, pp. 495–520, 2016.

[305] A. Salovaara, K. Lyytinen, and E. Penttinen, "High reliability in digital organizing: Mindlessness, the frame problem, and digital operations", *MIS quarterly*, vol. 43, no. 2, pp. 555–578, 2019.

[306] G. Wang, O. Henfridsson, J. Nandhakumar, and Y. Yoo, "Product meaning in digital product innovation.", *MIS quarterly*, vol. 46, no. 2, pp. 947–976, 2022.

[307] R. Yin, *Case Study Research* (Applied social research methods series). SAGE Publications, 2014, ISBN: 9781452242569. [Online]. Available: https://books.google.is/books?id=Cdk5DQAAQBAJ.

[308] M. D. Myers and M. Newman, "The qualitative interview in is research: Examining the craft", *Information and organization*, vol. 17, no. 1, pp. 2–26, 2007.

[309] U. Schultze and M. Avital, "Designing interviews to generate rich data for information systems research", *Information and organization*, vol. 21, no. 1, pp. 1–16, 2011.

[310] T. Speith, "A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods", in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '22, New York, NY, USA: Association for Computing Machinery, Jun. 2022, pp. 2239–2250, ISBN: 978-1-4503-9352-2. DOI: 10.1145/3531146.3534639. [Online]. Available: https://dl.acm.org/doi/10.1145/3531146.3534639.

[311] S. Najdi, A. A. Gharbali, and J. M. Fonseca, "Feature ranking and rank aggregation for automatic sleep stage classification: A comparative study", *BioMedical Engineering OnLine*, vol. 16, no. 1, p. 78, Aug. 2017, ISSN: 1475-925X. DOI: 10.1186/s12938-017-0358-3. [Online]. Available: https://doi.org/10.1186/s12938-017-0358-3.

[312] V. Braun and V. Clarke, "Using thematic analysis in psychology", *Qualitative research in psychology*, vol. 3, no. 2, pp. 77–101, 2006.

[313] A. Dubois and L.-E. Gadde, "Systematic combining: An abductive approach to case research", *Journal of Business Research*, Markets as Networks, vol. 55, no. 7, pp. 553–560, Jul. 2002, ISSN: 0148-2963. DOI: 10.1016/S0148-2963(00)00195-8. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0148296300001958.

[314] J. S. Ancker, A. Edwards, S. Nosal, *et al.*, "Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system", eng, *BMC medical informatics and decision making*, vol. 17, no. 1, p. 36, Apr. 2017, ISSN: 1472-6947. DOI: 10.1186/s12911-017-0430-8.

[315] F. Cabitza, R. Rasoini, and G. F. Gensini, "Unintended Consequences of Machine Learning in Medicine", eng, *JAMA*, vol. 318, no. 6, pp. 517–518, Aug. 2017, ISSN: 1538-3598. DOI: 10.1001/jama.2017.7797.

[316] N. Siddiqi, Ed., *Intelligent credit scoring: building and implementing better credit risk scorecards*, eng, 2nd edition. Hoboken, New Jersey: Wiley, 2017, ISBN: 978-1-119-28229-7 978-1-119-28239-6.

[317] C. Bravo, S. Maldonado, and M. Óskarsdóttir, *Deep Learning in Banking: Integrating Artificial Intelligence for Next-Generation Financial Services*, 1st ed. Hoboken: John Wiley & Sons, Inc, 2025, ISBN: 978-1-394-29537-1.

[318] Alþingi, *Lög um fasteignalán til neytenda*, 2016. [Online]. Available: https://www.althingi.is/lagas/nuna/2016118.html.

[319] *Reglugerð um fasteignalán til neytenda*, 2017. [Online]. Available: https://island.is/reglugerdir/nr/0270-2017.

[320] S. Íslands, *Reglur um hámark veðsetningarhlutfalls fasteignalána til neytenda.* 2024.

Appendix A

Paper IV supplement

A.1 Bland–Altman analysis

To characterize the distribution of scorer-versus-panel differences at the metric level, we performed Bland–Altman analyses using the same leave-one-out panel reference defined above. For each recording i, scorer A, and sleep metric π, let

$$Y_{\pi,i}^{(A)}$$

denote the scorer-derived metric value and let

$$\bar{Y}_{\pi,i}^{(-A)} = \frac{1}{K-1} \sum_{\substack{B=1 \\ B \neq A}}^{K} Y_{\pi,i}^{(B)}$$

denote the corresponding leave-one-out panel mean across the other $K - 1 = 9$ scorers.

For each scorer-recording pair, the Bland-Altman signed difference was defined as

$$d_{\pi,i}^{(A)} = Y_{\pi,i}^{(A)} - \bar{Y}_{\pi,i}^{(-A)},$$

and the corresponding Bland–Altman mean was defined as

$$m_{\pi,i}^{(A)} = \frac{Y_{\pi,i}^{(A)} + \bar{Y}_{\pi,i}^{(-A)}}{2}.$$

For each metric, pooled Bland–Altman summary statistics were then computed across all scorer-recording pairs. Let

$$\bar{d}_\pi = \frac{1}{NK} \sum_{i=1}^{N} \sum_{A=1}^{K} d_{\pi,i}^{(A)}$$

denote the pooled mean difference and let

$$s_{d,\pi} = \mathrm{sd}_{i,A}\left(d_{\pi,i}^{(A)}\right)$$

denote the pooled standard deviation of the signed differences. The 95% limits of agreement (LoA) were computed as

$$\mathrm{LoA}_\pi^{\mathrm{lower}} = \bar{d}_\pi - 1.96\, s_{d,\pi}, \qquad \mathrm{LoA}_\pi^{\mathrm{upper}} = \bar{d}_\pi + 1.96\, s_{d,\pi}.$$

Because the reference for each scorer is the mean of the remaining scorers, the pooled mean difference $\bar{d}_\pi$ is zero by construction when aggregated across all scorers and recordings. Accordingly, Table A.1 reports the pooled standard deviation of the signed differences together with the corresponding symmetric lower and upper 95% limits of agreement. Figure A.1 shows Bland–Altman plots for selected metrics, with each point representing one scorer-recording comparison.

Bland–Altman analyses showed tight agreement for SE, REM%, and SOL, but wider spread and evidence of heteroscedasticity for several NREM stage metrics and for ROL.

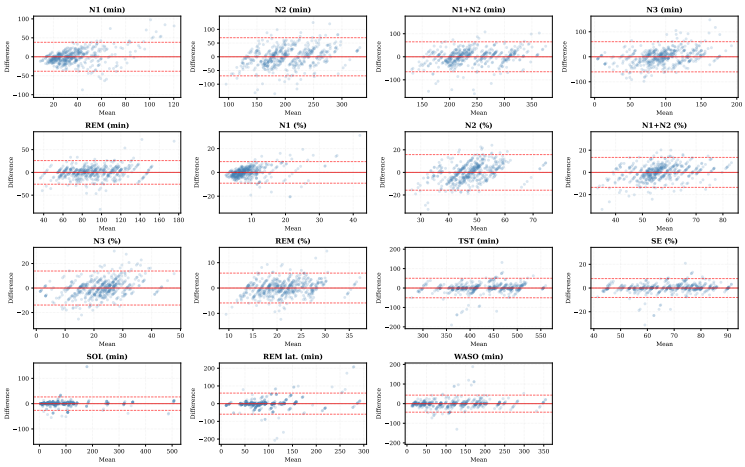


Figure A.1: Bland-Altman plots for selected derived sleep metrics using scorer-versus-leave-one-out consensus comparisons. Each point represents one scorer–recording comparison. The solid red line denotes the pooled mean difference, and the dashed orange lines denote the 95% limits of agreement. The tightest agreement was observed for SE and REM%, whereas SOL showed broader limits of agreement; broader spread and larger outliers were observed for several NREM metrics and for REM onset latency.

Table A.1: Pooled Bland–Altman summary statistics from scorer-versus-leave-one-out comparisons.

Metric	SD of differences	Lower 95% LoA	Upper 95% LoA
TST	25.63	-50.23	50.23
N1 duration	19.49	-38.20	38.20
N2 duration	35.48	-69.53	69.53
N1/N2 duration	33.10	-64.88	64.88
N3 duration	30.78	-60.34	60.34
REM duration	13.25	-25.97	25.97
N1%	4.63	-9.07	9.07
N2%	8.05	-15.77	15.77
N1/N2%	6.89	-13.50	13.50
N3%	7.08	-13.88	13.88
REM%	3.00	-5.88	5.88
SOL	13.56	-26.58	26.58
ROL	30.41	-59.61	59.61
WASO	22.17	-43.45	43.45
SE	4.06	-7.96	7.96

LoA denotes limits of agreement.
Abbreviations: SD, standard deviation; TST, total sleep time; N1, stage N1 sleep; N2, stage N2 sleep; N3, stage N3 sleep; REM, rapid eye movement; SOL, sleep onset latency; ROL, rapid eye movement onset latency; WASO, wake after sleep onset; SE, sleep efficiency.

A.2 Pairwise sleep-metric disagreement

A.2.1 Pairwise scorer-versus-scorer sleep-metric disagreement

To characterize direct disagreement between human scorers, we analyzed all pairwise scorer combinations within each recording. For recording i and scorer pair (A, B), the absolute pairwise disagreement for sleep metric π was defined as

$$d_{\pi,i}^{(A,B)} = \left| Y_{\pi,i}^{(A)} - Y_{\pi,i}^{(B)} \right|,$$

where

$$Y_{\pi,i}^{(A)} = \Psi_\pi\left(H_i^{(A)}\right)$$

denotes the value of metric π derived from scorer A's hypnogram for recording i.

Pairwise disagreement was computed for the full set of derived sleep metrics analyzed in the study. These absolute pairwise differences were then summarized across all scorer pairs and recordings using descriptive statistics, including the mean, median, first quartile, third quartile, and maximum (Table A.2).

Table A.2: Summary of pairwise absolute disagreement for derived sleep metrics across all scorer pairs and recordings.

Sleep metric	Mean abs. diff.	Median abs.diff.	Q1	Q3	Max
TST	18.56	9.00	3.50	21.00	244.50
N1 duration	18.44	13.50	6.00	24.50	154.50
N2 duration	36.19	28.50	13.00	51.00	230.50
N1/N2 duration	31.96	24.25	10.50	43.00	233.00
N3 duration	30.61	23.00	10.00	43.50	207.00
REM duration	12.30	8.50	3.50	16.50	107.00
N1%	4.28	3.13	1.45	5.53	46.51
N2%	8.42	6.86	3.28	11.85	51.13
N1/N2%	6.91	5.25	2.42	9.67	45.93
N3%	7.05	5.29	2.27	9.72	39.15
REM%	2.99	2.27	1.02	4.17	20.09
SOL	6.59	1.50	0.50	5.50	178.00
ROL	16.87	3.00	1.00	11.00	374.50
WASO	16.17	8.50	3.50	19.00	235.00
SE	2.93	1.43	0.59	3.37	40.05

Values summarize absolute pairwise disagreement across all scorer pairs and recordings.