Sleep Stage Classification from Wearable Sensor Data: Toward Transparent and Accessible Health AI

Brandon Ismalej and Dr. Xunfei Jiang
Department of Computer Science
California State University, Northridge

OFFICE OF UNDERGRADUATE RESEARCH

ARCS AUTONOMY RESEARCH CENTER FOR STEAHM

NASA Partner

Scan to explore the full project

Introduction

Context & Motivation:

- **Essential Role of Sleep:** Supports cognitive function, physical health, and overall well-being
- **Clinical Limitations:** Polysomnography is accurate but expensive and inconvenient (labor-intensive, clinical setting)
- **Wearable Potential:** Smartwatches and non-invasive sensors offer continuous, real-world data

Goals & Objectives:

- **Develop Explainable Machine Learning Models:** Classify sleep stages from wearable signals with known algorithm and transparent data use
- **Improvable Accessibility:** Provide a cost-effective, user-friendly approach to personalized health insights through an open-source platform

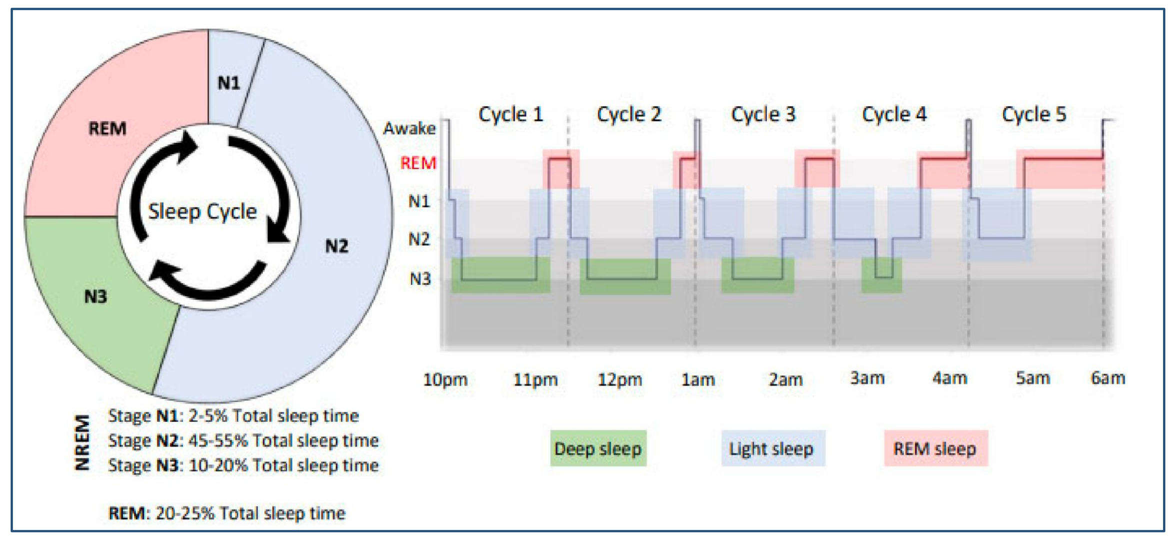
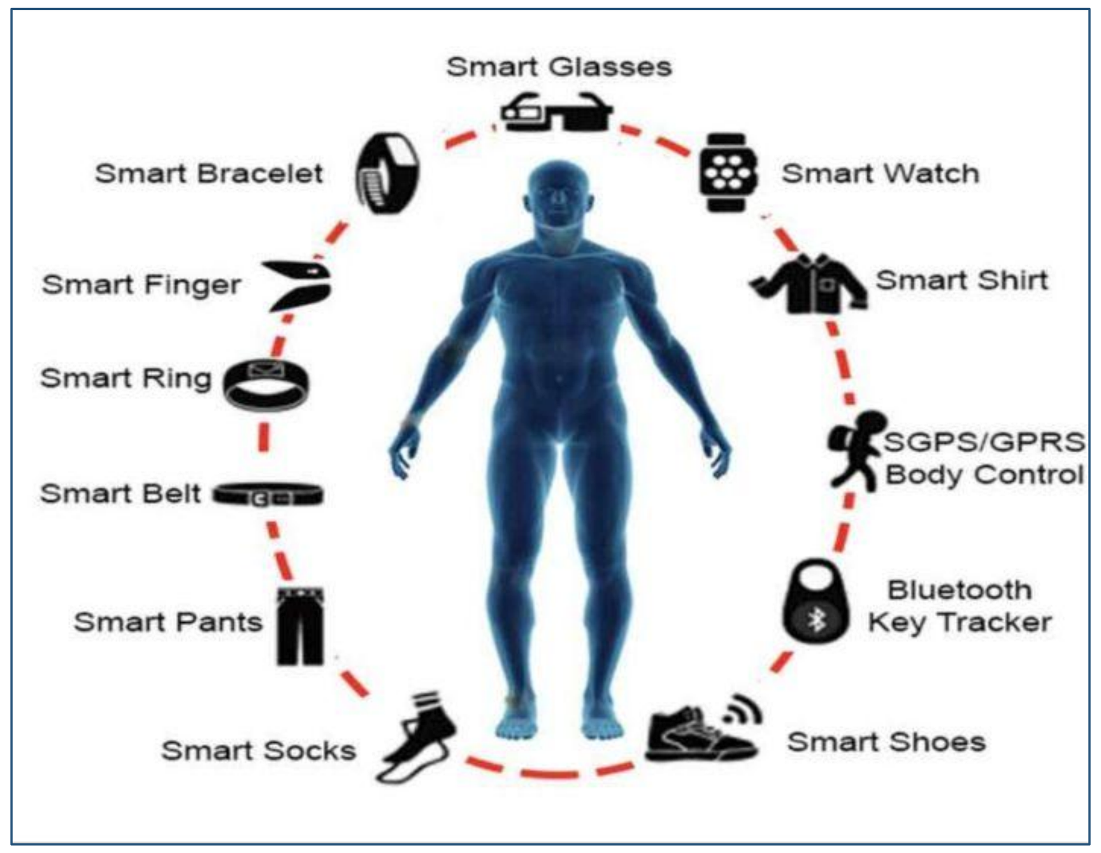


Figure 1: Hypnogram Depicting Sleep Stage Distribution Across Cycles [1]



BORACLE:

- **Boracle Platform** [4]**:** This project contributes to Boracle, an open-source cyber-physical system for collecting, standardizing, and analyzing health data from wearables.
- **Transparent Health AI:** Aligned with Boracle's mission to make ML-driven health insights more accessible and understandable by enabling algorithm transparency and user control.

Methodology

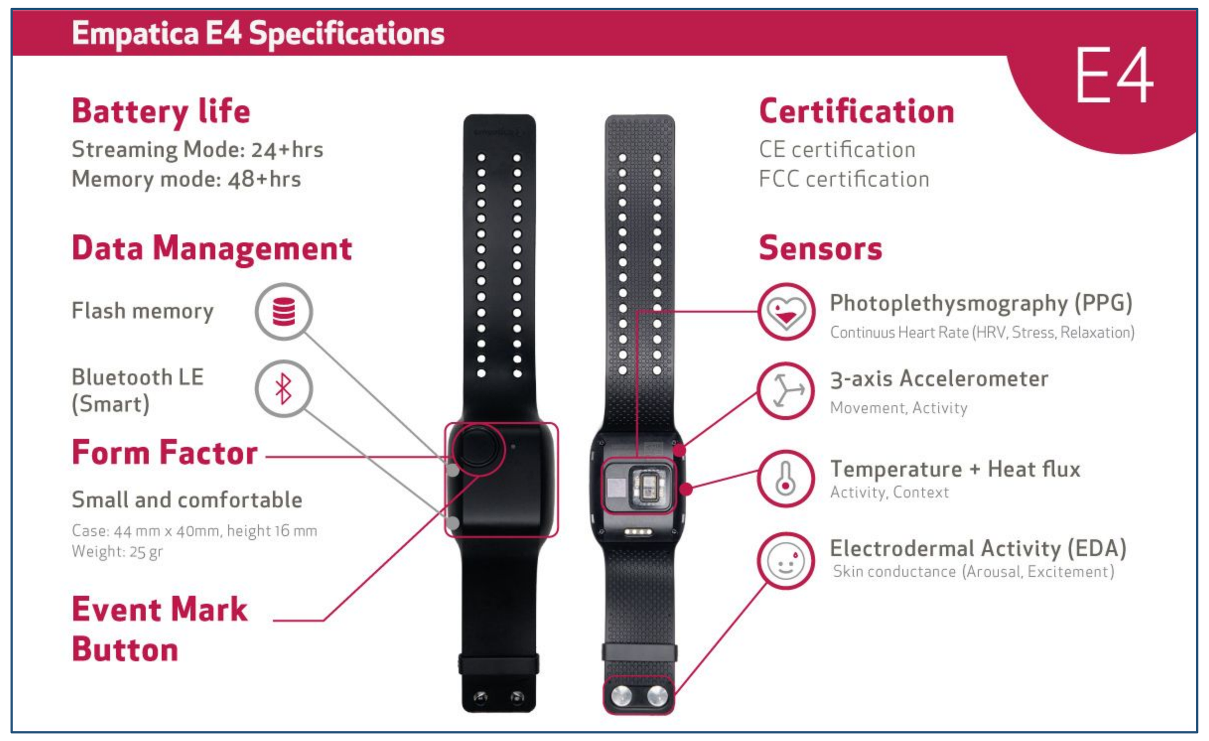


Figure 2: Empatica E4 Wristband with Core Specifications [2]

Dataset Overview:

- **DREAMT:** Dataset for Real-time sleep stage EstimAtion using Multisensor wearable Technology [3]
  - **100 individuals** recruited from Duke University Health System Sleep Disorder Lab
  - **Wearable (Empatica E4) signals:**
    - Raw: Blood Volume Pulse (BVP), Accelerometry (ACC_X,Y,Z), Electrodermal Activity (EDA), Skin Temperature (TEMP)
    - Derived: Heart Rate (HR), Inter-Beat Interval (IBI)
    - Sleep Stage Labels: Technician-annotated (PSG) every 30 seconds (W, N1, N2, N3, R)
- Chosen for its high-accuracy, sensor availability and signal collection capability

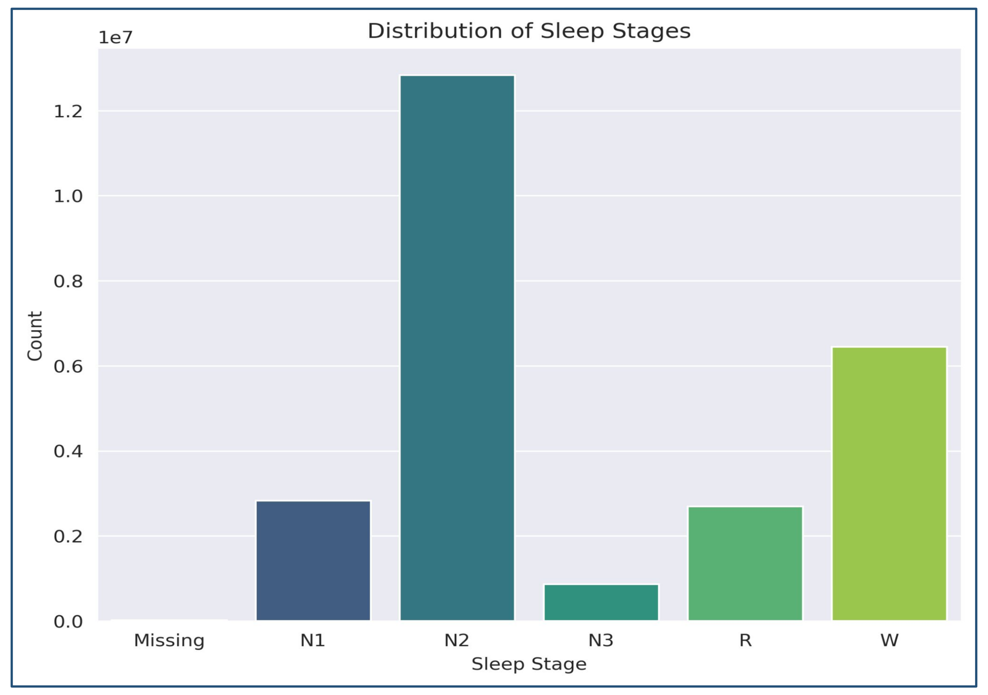


Figure 3: Distribution of Sleep Stage Frequency

Exploratory Data Analysis (EDA) Insights:

- Sufficient Data Per Participant: ~7 hours of data/participant
- Natural Class Imbalance: N2 is the most frequent stage, N3 is underrepresented, potentially impacting model performance

Data Preprocessing:

- **Filter** out preparation (P) stages
- **Raw signals** collected at 64 HZ:
  - Downsample to 10 Hz for full model
  - Downsample to 1 Hz for reduced model
- 4th-order **Butterworth low-pass filter** to BVP, ACC, EDA, and TEMP **to remove high-frequency noise**
- **Majority voting** with smoothing over downsampled windows to **maintain sleep stage transitions**

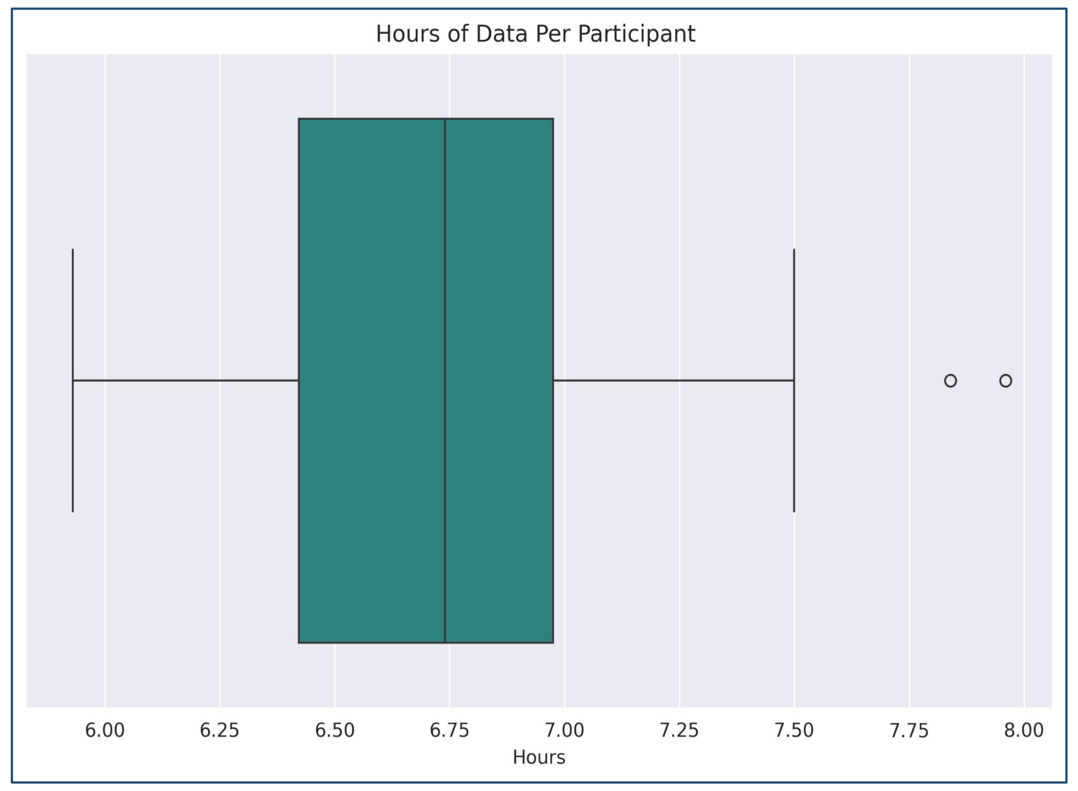


Figure 4: Hours of Data Per Participant

Neural Network Development & Training:

- **Neural Network Architectures**
  - Long-Short Term Memory (LSTM)
  - Convolutional Neural Network (CNN)
- **Training Setup:**
  - Train-Test Split of Data: 70% train, 15% validation, 15% test
  - Class Imbalance Mitigation
    - Focal Loss
    - Class weighting
  - **Full-Feature Models:**
    - BVP, ACC_X,Y,Z, EDA, TEMP, HR, IBI
  - **Reduced-Feature Models:**
    - HR, IBI, ACC_X,Y,Z

Results

Full-Feature Model:

- LSTM: Sequences of 10 seconds at 10 Hz
- CNN:Sequences of 10 seconds at 10 Hz

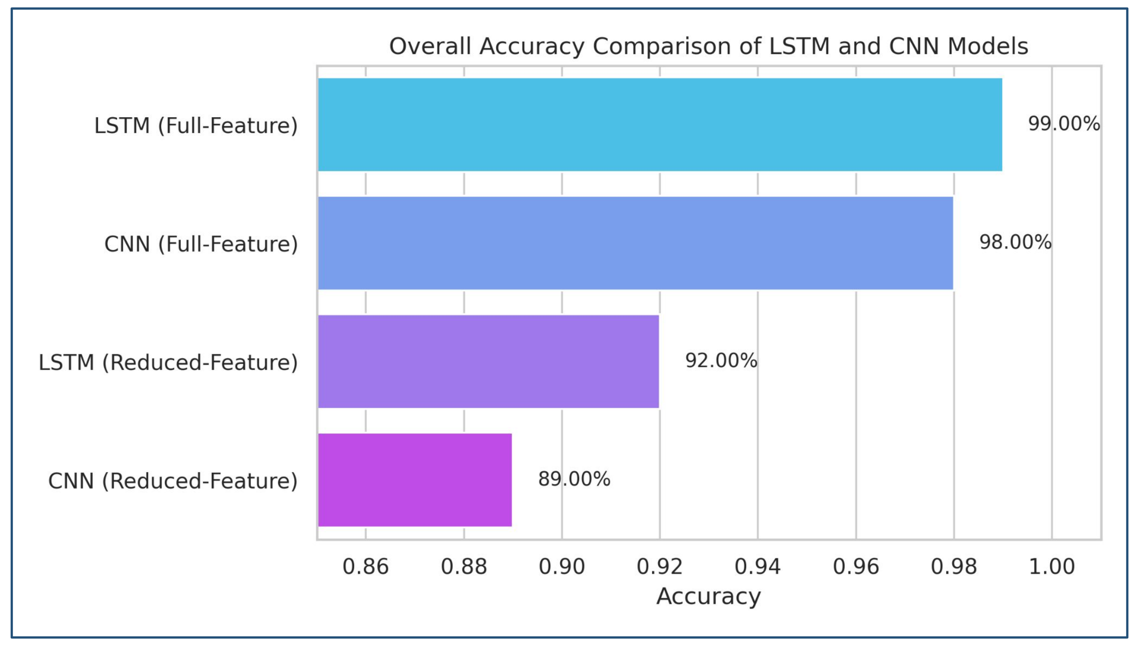


Figure 4: Overall Accuracy Comparison of LSTM & CNN

Reduced-Feature Model:

- LSTM: Sequences of 30 seconds at 1 Hz
- CNN: Sequences of 30 seconds at 1 Hz

	Precision	Recall	F1-Score
Wake	0.99	0.99	0.99
N1	0.98	0.99	0.98
N2	1.00	1.00	1.00
N3	0.90	0.96	0.93
REM	1.00	1.00	1.00

	Precision	Recall	F1-Score
Wake	0.99	0.93	0.96
N1	0.88	0.89	0.88
N2	0.93	0.93	0.93
N3	0.59	0.84	0.70
REM	0.95	0.91	0.93

Table 1, 2: Classification Report of LSTM; Full-Feature (left), Reduced Feature (right)

	Precision	Recall	F1-Score
Wake	0.99	0.98	0.98
N1	0.91	0.93	0.92
N2	0.97	0.98	0.98
N3	0.72	0.83	0.77
REM	0.96	0.96	0.96

	Precision	Recall	F1-Score
Wake	0.96	0.91	0.94
N1	0.83	0.87	0.85
N2	0.91	0.93	0.92
N3	0.75	0.81	0.78
REM	0.94	0.92	0.93

Table 3, 4: Classification Report of CNN; Full-Feature (left), Reduced Feature (right)

Conclusion & Future Work

- Our methodology effectively classifies sleep stages using smartwatch sensor data, achieving high classification performance with few, accessible features

Future Work:

- Explore domain adaptation techniques to improve robustness across different wearable devices
- Deployment of this model to the Boracle Platform for use with user wearable devices

References

[1] M. W. Driller *et al.*, "Pyjamas, Polysomnography and Professional Athletes: The Role of Sleep Tracking Technology in Sport," *Sports*, vol. 11, no. 1, Art. no. 1, Jan. 2023, doi: 10.3390/sports11010014.
[2] "E4 wristband technical specifications," *Empatica Support*, 2025. https://support.empatica.com/hc/en-us/articles/202581999-E4-wristband-technical-specifications (accessed Mar. 11, 2025).
[3] K. Wang, J. Yang, A. Shetty, and J. Dunn, "DREAMT: Dataset for Real-time sleep stage EstimAtion using Multisensor wearable Technology," Physionet.org, Feb. 05, 2025. https://physionet.org/content/dreamt/2.0.0/ (accessed Mar. 26, 2025).
[4] C. A. Dimalanta, M. Smith, K. Bonakdar, X. Jiang, N. Ho, and T. Dung, "Poster - Boracle: An Open Data Platform For Health Condition Prognostics," 2024 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE), pp. 200–201, Jun. 2024, doi: https://doi.org/10.1109/chase60773.2024.00038.