Sleep Stage Classification from PPG-Derived Heart Rate and Wrist Motion

Ch. Muhammad Ahsan
Department of Data Science
Faculty of Computing and Information Technology
University of the Punjab
Lahore, Pakistan
bsdsf23a021@pucit.edu.pk

Taha Saleem
Department of Data Science
Faculty of Computing and Information Technology
University of the Punjab
Lahore, Pakistan
bsdsf23a043@pucit.edu.pk

Fahad Rahman
Department of Data Science
Faculty of Computing and Information Technology
University of the Punjab
Lahore, Pakistan
bsdsf23a024@pucit.edu.pk

Abdul Rahman Tahir
Department of Data Science
Faculty of Computing and Information Technology
University of the Punjab
Lahore, Pakistan
bsdsf23a038@pucit.edu.pk

Faisal Bukhari
Department of Data Science
Faculty of Computing and Information Technology
University of the Punjab
Lahore, Pakistan
faisal.bukhari@pucit.edu.pk

***Abstract*—Polysomnography (PSG) is the clinical gold standard for sleep staging but is costly and impractical for continuous, at-home monitoring. Wearable devices provide an attractive alternative by passively recording photoplethysmography (PPG)-derived heart rate and wrist motion; however, these signals are noisy, indirect proxies of neural sleep dynamics, and must be processed under tight computational constraints. This work presents a lightweight sleep staging pipeline using the Sleep-Accel dataset (Apple Watch heart rate, tri-axial acceleration, step counts, and PSG-scored labels). We extract per-epoch physiological and motion summaries, augment them with multi-scale temporal context, and train a histogram-based gradient boosting classifier with subject-wise evaluation. To enforce plausible stage continuity, we apply Hidden Markov Model (HMM) smoothing with Viterbi decoding using transition statistics learned from training subjects. On a held-out subject test split, we obtain accuracy of 0.526/0.560/0.707 for 5-/4-/3-stage classification, respectively, demonstrating an efficient and interpretable approach suitable for wearable deployment.**

***Index Terms*—Sleep staging, wearables, photoplethysmography, actigraphy, gradient boosting, Hidden Markov Model, temporal context.**

I. Introduction

Sleeping is an important function that helps us remain physically and mentally sound. Understanding how we transition between levels of sleep helps identify any issue with insomnia, apnea, or quality of sleep [1], [2]. Currently, the standard of care in this area is polysomnography (PSG), which employs signal modalities such as EEG, EOG, and EMG. However, PSG tests are quite expensive, cumbersome and not very practical for use in everyday life [2], [3].

In recent years, wearable devices have become an easier and more affordable way to collect sleep-related data. These gadgets can record signals such as heart rate derived from photoplethysmography (PPG) and wrist motion from accelerometers, and some devices also provide blood oxygen saturation and skin temperature [2], [4], [5]. Still, they face some real challenges. Data from wearables is often noisy, and the sensors do not directly capture brain activity like EEG does. On top of that, wearables have limited power, memory, and processing capacity [6]–[8].

For addressing such problems, various techniques of machine learning and deep learning have been investigated to detect the stages of sleep automatically. However, models with high accuracy, like those implemented via deep models including ”convolutional neural networks (CNNs)” and ”long short-term memory (LSTM)” models, are computationally intensive, meaning they are not suited for usage in devices like those mentioned above [1], [3], [9], [10]. However, simpler models process faster, but may lack accuracy at times. However, considerable progress has been made in integrating predictors for sleep stages that are light in computation but still model the transitions between the stages of sleep in a probabilistic manner, which is efficient enough for execution on wearables [1], [5], [11], [12].

This paper presents a study on the classification of sleep

stage using wearable-friendly signals, namely *PPG-derived heart rate* and *wrist motion*. The idea is to reach the right compromise between performance and efficiency in order to enable effective sleep staging without consuming device resources. For better temporal reliability and purposes of interpretation, we add a probabilistic smoothing component regarding plausible processes of the sleep stages [2], [4], [6], [13].

II. Dataset and Preprocessing

A. Dataset Description

In particular, we utilize the PhysioNet Sleep-Accelel dataset, which provides tri-axial wrist acceleration - units of g, PPG-derived heart rate - bpm, and step counts recorded from Apple Watch devices during overnight PSG sessions, and PSG-scored sleep stage labels [14], [15].

B. Epoching and Label Mapping

All time series were aligned into non-overlapping 30-second epochs. In the case of each subject, the label file provides time stamps and stage codes. Following the convention of the data set, we map stage codes $\{0, 1, 2, 3, 5\}$ to the 5-stage label set (Wake, N1, N2, N3, REM by mapping code 5 to REM. We also report two widely used coarser resolutions:

- **4-stage**: Wake, Light (N1+N2), Deep (N3), REM
- **3-stage**: Wake, NREM (N1+N2+N3), REM

C. Wearable Feature Extraction

For the computation to remain light and adaptable to the limitations posed by wearables, we calculated the summary stats for each epoch with regard to motion, heart rate, and steps taken.

1) Motion Features: For acceleration samples (x_t, y_t, z_t) within an epoch, we determined the vector magnitude

$$v_t = \sqrt{x_t^2 + y_t^2 + z_t^2}, \tag{1}$$

and derived low-cost estimates like mean and standard deviation of v_t, maximum v_t, and means and standard deviations along the axes. Additionally, we calculated the ENMO (Euclidean Norm Minus One) [16]:

$$\text{ENMO}_t = \max(v_t - 1, 0), \tag{2}$$

and used the per-epoch mean ENMO. Lastly, we measured the number of motion samples per epoch as a simple indicator of quality/coverage.

2) Heart Rate and HRV-Proxy Features: In each epoch and for samples of heart rate, we calculated mean, standard deviation, minimum, and maximum heart rate. Because raw PPG waveform data are not provided, from each heart rate measurement, a surrogate IBI metric can be constructed as follows:

$$\text{IBI} = \frac{60}{\text{HR}}. \tag{3}$$

In each epoch, we estimated the mean and standard deviation of the IBI, as well as RMSSD-like and mean absolute difference between IBIs measures, which represent short-term variability. Also, the number of samples of the HR in each epoch was stored.

3) Steps and Time-of-Night Features: We aggregated step count by epoch and, when available, simple summaries such as min, max, mean, and standard deviation of the step count. In addition, we represented time of night with a feature $t \in [0, 1]$, given by epoch index divided by total number of epochs, encoded with sinusoidal functions:

$$\sin(2\pi t),\ \cos(2\pi t), \tag{4}$$

to capture position-in-recording effects.

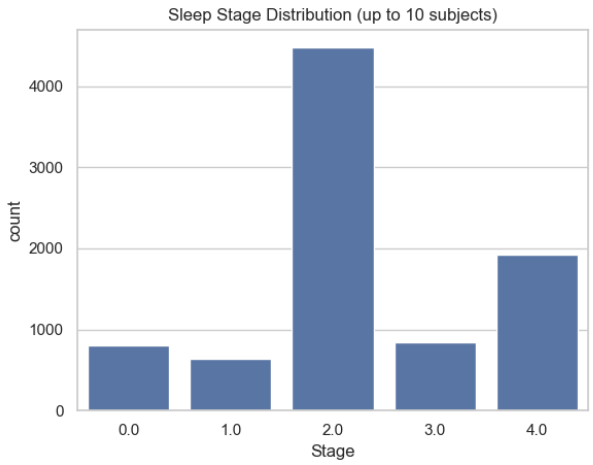


Fig. 1. Sleep stage distribution across the analyzed epochs, illustrating pronounced class imbalance (notably minority N1).

D. Missing Data Handling

The missing heart rate episodes were represented by using a binary missingness matrix. The heart rate means were imputed by within-subjects interpolation and fallback median imputation, and the other variables by conservative imputation (e.g., stepping imputation = 0 when missing).

III. Exploratory Statistical Analysis

A. Class Distribution

In Fig. 1, the distribution of our experimental epochs with the five stages of sleeping can be viewed. It is noteworthy that there are more occurrences in stage N2, while stage N1 stages fewer occurrences. These consequences imply an interpretation mechanism for our classification results, including weighting of classes.

B. Correlation Structure of Wearable Features

We calculated the engineered feature redundancy by computing the pairwise correlations. Fig. 2 (Spearman) shows that the engineered features are strongly correlated within groups, i.e., accelerometer-derived summaries with each other, heart-rate summaries with each other, and step-derived summaries with each other. Cross-sensor modality pairs, e.g., HR vs.

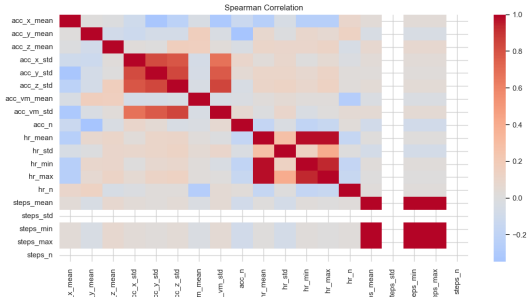


Fig. 2. Spearman correlation heatmap among motion, heart-rate, and step-derived features. Strong within-modality correlation suggests redundancy, while weaker cross-modality correlation suggests complementary information.

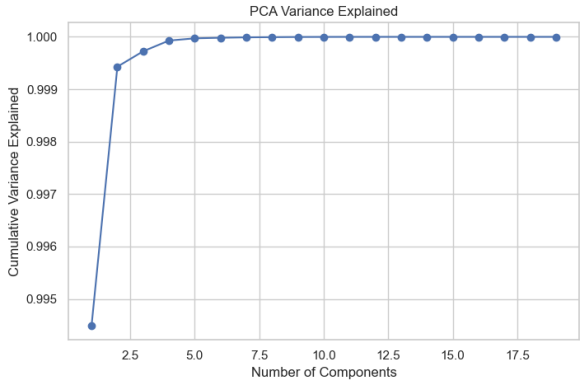


Fig. 3. Cumulative explained variance by PCA components. A small number of components accounts for nearly all variance, indicating substantial redundancy among engineered features.

motion show generally weaker correlations suggesting complementary information across sensor modalities. The similar pattern from Pearson/Spearman for most pairs suggests relationships that are predominantly monotonic and close to linear [17].

C. Principal Component Analysis (PCA)

We applied PCA as an exploratory check on effective dimensionality [18]. Fig. 3 illustrates that only a very few principalcomponents explain almost all the variability appearing in the sensor set after applying the standardized feature transform, which verifies the correlation blocks discovered in the previous experiment. Interpretability of results, together with ease of deployment, means that we use the original set of engineered features, which are not mapped to any hidden representation, with the only purpose of applying the PCA as a diagnostic test.

D. Feature Discriminability Across Stages (ANOVA and Effect Sizes)

To quantify whether feature means differ across sleep stages, we performed one-way ANOVA per feature [19]. Table I summarizes representative results, including effect sizes (η^2 and ω^2). Motion-variability features (e.g., axis standard deviations

TABLE I
ONE-WAY ANOVA FEATURE RANKING ACROSS 5 SLEEP STAGES (SELECTED FEATURES). LARGER η^2 / ω^2 INDICATE STRONGER STAGE SEPARATION.

Feature	F	p-value	η^2	ω^2
steps_min	2.666	0.0327	0.00030	-0.00015
acc_y_mean	4.972	0.00053	0.00055	0.00011
acc_z_mean	17.966	≈ 0	0.00199	0.00155
acc_x_mean	25.109	≈ 0	0.00278	0.00234
acc_vm_mean	36.877	≈ 0	0.00407	0.00363
hr_n	52.924	≈ 0	0.00584	0.00540
hr_mean	69.361	≈ 0	0.00764	0.00720
hr_max	74.148	≈ 0	0.00816	0.00772
hr_min	87.561	≈ 0	0.00962	0.00918
acc_n	140.931	≈ 0	0.01539	0.01496
hr_std	227.275	≈ 0	0.02459	0.02416
acc_z_std	351.499	≈ 0	0.03753	0.03710
acc_vm_std	387.775	≈ 0	0.04125	0.04082
acc_y_std	392.345	≈ 0	0.04171	0.04129
acc_x_std	396.495	≈ 0	0.04213	0.04171

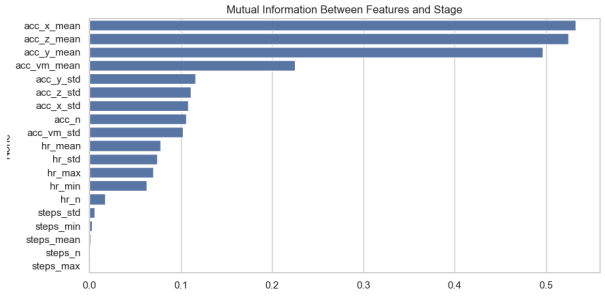


Fig. 4. Mutual information ranking of engineered features. Motion-derived summaries dominate, while step summaries contribute minimally.

and vector-magnitude standard deviation) show the largest effects, followed by heart-rate variability, while step summaries contribute weakly. These are in line with the expectation that wake stages involve higher and sharper activity patterns, with increasing stability during the deeper levels of sleep

We also investigated pairwise standardized mean differences (Cohen's d) as an interpretable measure of effect size [20]. The biggest differences lie between Wake and stable sleep (particularly N2) with very large d for lateral motion variability. Many measures have effect sizes close to zero between N2/N3/REM, which represents the challenging nature of precise staging based solely on HR and motion.

E. Mutual Information Feature Importance

As a model-agnostic relevance check, we computed mutual information (MI) between each feature and the 5-stage labels [21]. Fig. 4 indicates that accelerometer mean/variability summaries are most informative, while step-derived summaries contribute little.

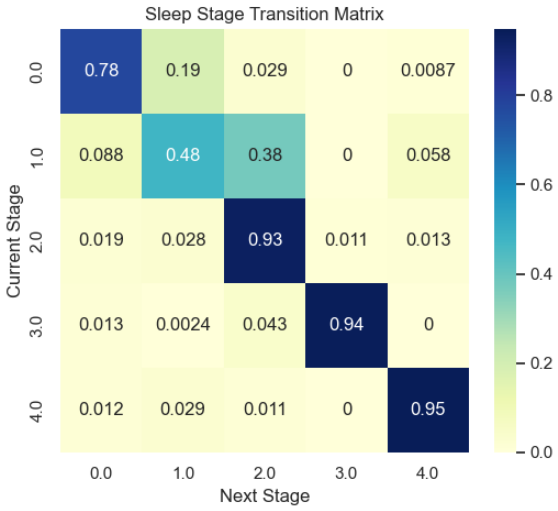


Fig. 5. Empirical sleep-stage transition probability matrix across consecutive epochs. Strong diagonal dominance indicates high temporal persistence.

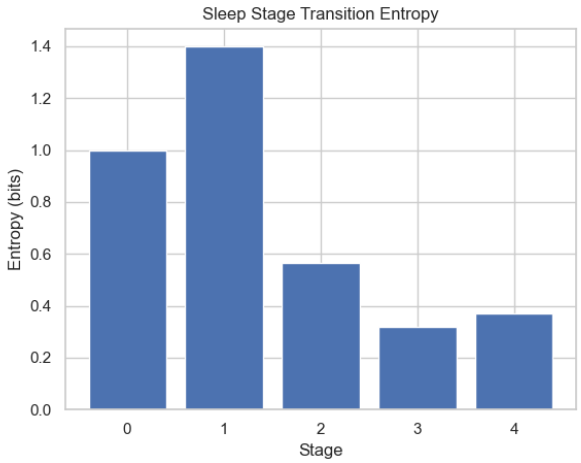


Fig. 6. Transition entropy per stage: N1 is most unpredictable, while deeper stages are more stable.

F. Sleep-Stage Temporal Dynamics

We computed the empirical transition matrix between consecutive 30-second epochs. Fig. 5 shows high persistence on the diagonal and structured transitions consistent with sleep onset dynamics.

We also computed transition entropy (Shannon entropy) per stage [22]. Fig. 6 shows that N1 has the highest entropy, while deeper stages are more stable.

IV. Machine Learning Method

A. Temporal Context Augmentation

Sleep stage prediction requires temporal continuity and multi-minute trends. For each participant, we calculated rolling window statistics (mean and standard deviation) with window sizes of $w \in \{5, 10, 20, 40\}$ epochs, or 2.5 to 20 minutes, to capture multi-scale context information, and short-minus-long window differences for the interested variables.

TABLE II
Performance on Held-Out Subjects (Test Set: 5,781 epochs)

Task	Method	Acc.	Macro-F1	κ
5-stage	Raw	0.516	0.357	0.190
	HMM	0.526	0.350	0.189
4-stage	Raw	0.559	0.413	0.176
	HMM	0.560	0.410	0.171
3-stage	Raw	0.699	0.506	0.213
	HMM	0.707	0.510	0.221

B. Subject-Wise Train/Test Split

For preventing the leakage of subject information, an 80/20 subject-wise split is performed, which splits 80% of the subject data for training and the remaining 20% for testing. This split is used consistently throughout the experiments for 5-/4-/3-stage experiments.

C. Class-Imbalance Weighting

Sleep stages are imbalanced (especially N1). We used inverse-frequency class weights from the training set:

$$w_k \propto \left(\frac{N}{\max(n_k, 1)} \right)^{\gamma}, \tag{5}$$

with $\gamma = 1$, normalized to mean 1.

D. Gradient Boosting Classifier

We trained a histogram-based gradient boosting classifier on the engineered features for efficient learning on large tabular datasets. Key settings included depth-limited trees (max depth 12), learning rate 0.19, and early stopping with an internal validation fraction.

E. Probabilistic Sequence Smoothing via HMM

Motivated by strong stage persistence (Fig. 5), we applied HMM smoothing using Viterbi decoding [23], [24]. We estimated the transition matrix and initial distribution from training label sequences with additive smoothing; classifier predicted probabilities served as emissions.

V. Results

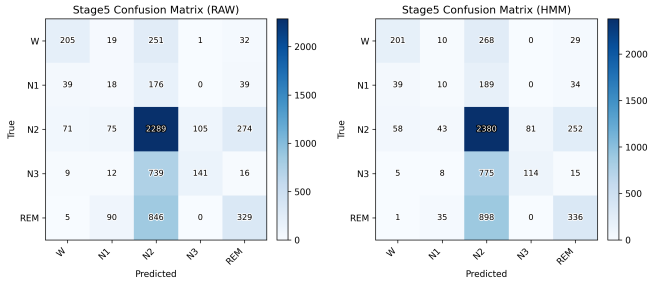
A. Evaluation Metrics

We report accuracy, macro-averaged F1, and Cohen's κ, along with per-class precision/recall/F1 and confusion matrices.

B. Classification Performance

Table II summarizes results for raw per-epoch predictions and after HMM smoothing on the held-out test set (5,781 epochs).

Overall accuracy increases as the label space is coarsened (from 5-stage to 3-stage). HMM smoothing provides modest gains in accuracy, most notably for 3-stage classification, consistent with improved temporal consistency. Fig. 7 highlights that most confusions occur among N1/N2/N3.



(a) Raw (b) HMM

Fig. 7. 5-stage confusion matrices on the held-out subjects. HMM smoothing slightly increases temporal consistency but can further concentrate predictions into dominant classes (notably N2).

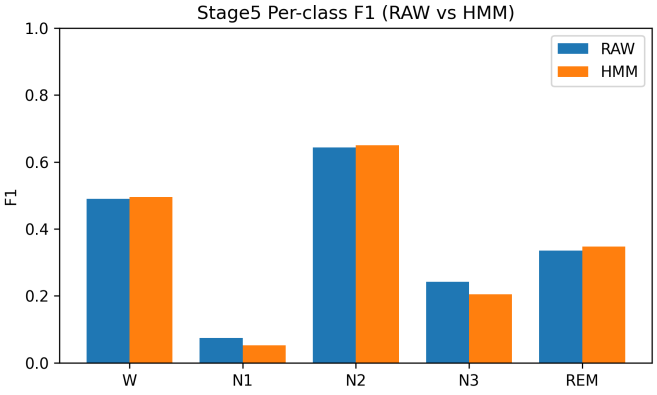


Fig. 8. Per-class F1 for 5-stage classification (Raw vs HMM). Minority classes (notably N1 and N3) remain the main failure modes.

C. Error Patterns

In 5-stage classification, the model strongly favors N2 (Fig. 7), reflecting both class imbalance and limited neurophysiological specificity in HR/motion signals. N1 is frequently confused with N2 and sometimes REM, and N3 is often mapped to N2 as well. Per-class F1 (Fig. 8) shows that HMM smoothing yields only minor changes and can trade minority sensitivity for smoother trajectories.

VI. Classification Diagnostics and Interpretability

A. Qualitative hypnogram example

Fig. 9 compares PSG labels against raw model predictions and HMM-smoothed predictions for one representative subject. The HMM reduces short-lived spikes and produces trajectories that better match expected stage persistence, even when the aggregate metric gains are modest.

B. Subject-wise impact of HMM smoothing

Overall metrics can hide subject-to-subject differences in wearable signal quality. Fig. 10 shows per-subject macro-F1 for RAW vs HMM (5-stage). Points above the diagonal indicate improvement with HMM; points below indicate cases where smoothing reduces minority-stage sensitivity.

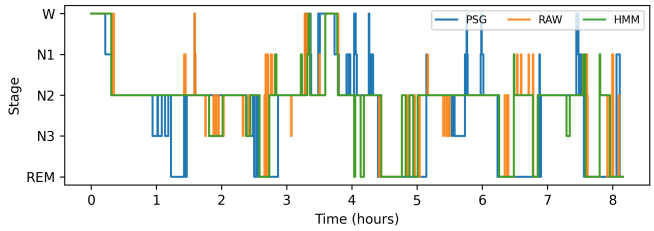


Fig. 9. Example hypnogram (5-stage): PSG ground truth vs RAW predictions vs HMM-smoothed sequence.

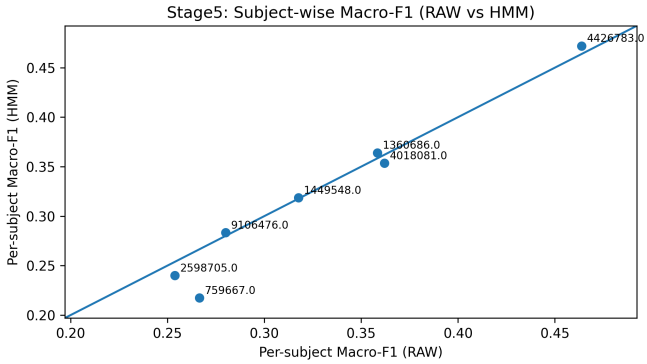


Fig. 10. Per-subject macro-F1 (5-stage): RAW vs HMM.

VII. Discussion

The proposed pipeline not only shows that wrist activity features and heart rate features from PPG have the potential to enable effective sleep stage analysis without resorting to large models, but it does so in an especially elegant and computationally efficient manner by exploiting gradient boosting on tabular features and HMM smoothing. The major limitation is the performance of the model at the minority stage, especially N1 and N3 (Figs. 7 and 8), hinting at the need for additional information (e.g., detailed HRV characteristics extracted from the raw signal of PPG, temperature, respiration variables, and modeling). On the whole, the biggest improvement comes from the use of multi-scale temporal context and the simulation of plausible stage duration conservation with HMM smoothing. More specifically, the main use of the variability-driven features is to better differentiate between the *Wake* and sleep stages, while the summaries and the HRV-proxies play more essential parts in distinguishing the overall sleep pattern during the night. Still, the distinction between the NREM stages N1, N2, and N3, when crossing the stages closely together, proves challenging based solely on the use of the HR and motion signals acquired by wearables.

VIII. Conclusion

We presented a wearable-friendly sleep staging approach using (i) lightweight epoch-level features from Apple Watch heart rate, accelerometry, and steps, (ii) multi-scale temporal context augmentation, (iii) a histogram-based gradient boosting classifier, and (iv) probabilistic HMM smoothing. Our

approach achieves 0.526 accuracy for 5-stage staging and 0.707 accuracy for 3-stage staging on held-out subjects and provides an efficient and interpretable baseline for on-device or near-device sleep analytics.

REFERENCES

[1] M. I. H. Siddiqui *et al.*, "Comparative analysis of traditional machine learning vs deep learning for sleep stage classification," May 2025, (preprint / conference paper): PDF on ResearchGate.

[2] S. Djanian, A. Bruun, and T. D. Nielsen, "Sleep classification using consumer sleep technologies and ai: A review of the current landscape," *Sleep Medicine*, 2022.

[3] M. Perslev, A. Darkner, J. M. Kempfner *et al.*, "U-sleep: resilient high-frequency sleep staging," *NPJ Digital Medicine*, 2021.

[4] V. Birrer, M. Perslev, A. J. Heidenreich *et al.*, "Evaluating reliability in wearable devices for sleep staging," *NPJ Digital Medicine*, 2024.

[5] S. K. H. S., "A systematic review of sensing technologies for wearable sleep staging," *PubMed*, 2021.

[6] J. Wang, Y. Guan, C. Chen, L. Zhou, L. T. Yang, and S. Gu, "On improving ppg-based sleep staging: A pilot study," *arXiv preprint*, Aug. 2025.

[7] A. Sathyanarayana, A. N. Goldstein, R. Haskell *et al.*, "Sleep quality prediction from wearable data using deep learning," 2016.

[8] M. Perslev *et al.*, "Benchmarking deep learning sleep staging systems and discussing deployment constraints," *NPJ Digital Medicine*, 2021.

[9] M. Perslev, P. Darkner, K. Jennum, and R. I. Kempfner, "U-sleep: a publicly available deep neural network for resilient sleep staging," *Nature (supplementary resources)*, 2021.

[10] D. Farooq and A. Author, "An active sleep monitoring framework using wearables," in *ACM Proceedings (Smart Health)*, (classic wearable sleep monitoring framework; useful for system design and feature engineering).

[11] H.-Y. Chih, T. Ahmed, A. P. Chiu *et al.*, "Multitask learning for automated sleep staging and wearable technology integration," *Advanced Intelligent Systems*, vol. 6, no. 1, Nov. 2023.

[12] U. Reimer and B. Emmenegger, "Recognizing sleep stages with wearable sensors in everyday settings," in *Proceedings of Smart Health / SmartCity Research (SCITEPRESS)*, 2017.

[13] L. van Hees and R. N. Jones, "Feature fusion and transition models for low-power sleep staging: survey and prospects," *IEEE Journal of Sensors*, 2022, (probabilistic/transition modeling background).

[14] O. Walch, Y. Huang, and D. Forger, "Sleep stage prediction with raw acceleration and photoplethysmography heart rate data derived from a consumer wearable device," *Sleep*, 2019.

[15] "Motion and heart rate from a wrist-worn wearable device and labeled sleep from polysomnography (sleep-accel) v1.0.0," PhysioNet, accessed 2025-12-29. [Online]. Available: https://physionet.org/content/sleep-accel/1.0.0/

[16] J. Olfermann, U. Ebner-Priemer, M. Reichert, and M. Giurgiu, "Comparability of accelerometry outcomes across popular metrics and widespread sensor positions," *PLOS ONE*, vol. 20, no. 5, p. e0324082, 2025.

[17] P. Schober, C. Boer, and L. A. Schwarte, "Correlation coefficients: Appropriate use and interpretation," *Anesthesia & Analgesia*, vol. 126, no. 5, pp. 1763–1768, 2018.

[18] M. Greenacre, P. J. F. Groenen, T. Hastie, A. Iodice D'Enza, A. Markos, and E. Tuzhilina, "Principal component analysis," *Nature Reviews Methods Primers*, vol. 2, p. 100, 2022.

[19] D. C. Montgomery, *Design and Analysis of Experiments*, 10th ed. Wiley, 2020.

[20] C. R. Brydges, "Effect size guidelines, sample size calculations, and statistical power in gerontology," *Innovation in Aging*, vol. 3, no. 4, p. igz036, 2019.

[21] Y. Polyanskiy and Y. Wu, *Information Theory: From Coding to Learning*, 1st ed. Cambridge University Press, 2025.

[22] ——, "Information theory: From coding to learning," *Cambridge University Press*, 2025, textbook.

[23] X. Ma, L. Luo, L. Liu, Z. Yan, X. Wang, and W. Yu, "Hidden markov models: Theory, algorithms, and applications in bioinformatics," *Genes & Diseases*, vol. 13, no. 1, p. 101729, 2025.

[24] ——, "Hidden markov models: Theory, algorithms, and applications in bioinformatics," *Genes & Diseases*, vol. 13, no. 1, p. 101729, 2025.