Sleep Stage Classification with Non-Linear Heart Rate Variability Methods and Deep Learning

Topi Niemi, Matias Kanniainen, Teemu Pukkila, Esko Toivonen, Esa Räsänen

Tampere University, Tampere, Finland

Abstract

We investigate sleep stage classification using heart rate variability (HRV), comparing conventional features—mean RR, root mean square of successive differences (RMSSD), low-frequency to high-frequency power ratio (LF/HF), SD1 and SD2 of the Poincaré plot—with those derived from dynamical detrended fluctuation analysis (DDFA). We trained three classifiers: logistic regression, eXtreme Gradient Boosting, and a neural network on each feature space independently, using physician-annotated sleep stages as ground truth, labeled as wake, rapid eye movement (REM) sleep, and non-rapid eye movement (NREM) sleep. The dataset comprised 2052 subjects without a history of sleep apnea, aged 40 to 89. DDFA consistently outperformed conventional HRV, improving NREM and REM classification across all models, indicating that it provides information beyond conventional HRV features for sleep stage classification. DDFA effectively captures correlations between consecutive heartbeats and are easily visualized and interpreted, allowing physiological links to different sleep stages. This interpretability makes DDFA features particularly valuable for explainable artificial intelligence, where understanding both features and model decisions is crucial.

1. Introduction

Sleep stages, including rapid eye movement (REM) and non-rapid eye movement (NREM; stages N1–N3), exhibit distinct physiological characteristics and typically cycle every 90 minutes, repeating 4–6 times per night. Accurate classification of these stages is clinically important, as many sleep disorders, such as sleep apnea and narcolepsy, are stage-dependent [1, 2]. Polysomnography (PSG) remains the gold standard for sleep assessment due to its comprehensive physiological monitoring [3], but its high cost and intrusive nature limit accessibility and scalability. At the same time, there is growing consumer interest in tracking sleep and recovery to optimize health, performance, and wellbeing, creating a market for wearable wellness applications and motivating the search for alternative, non-intrusive approaches to monitor sleep.

This increasing demand has fueled the use of physiological signals, such as heart rate variability (HRV), for sleep analysis. HRV is a well-established marker for assessing cardiac health, physical fitness, and sleep, and provides a non-invasive means to capture autonomic dynamics throughout the night. Conventional HRV (cHRV) metrics include, for example, the root mean square of successive RR interval (RRI) differences (RMSSD), the ratio of low- to high-frequency power (LF/HF), SD1 and SD2 from the Poincaré plot, and detrended fluctuation analysis (DFA). All of these features can be derived from wearable sensors and applied to sleep stage classification [4].

We focus on *dynamical* detrended fluctuation analysis (DDFA) [5], a recent extension of conventional DFA that enables simultaneous time- and scale-dependent examination of the scaling exponent in RRI sequences. We compare its performance with cHRV metrics in sleep stage classification. By applying identical machine learning architectures to different feature spaces, we specifically assess how the feature representation contributes to classification performance.

2. Data and preprocessing

We analyzed 5804 overnight PSG recordings from the first phase of the Sleep Heart Health Study (SHHS1), obtained via the National Sleep Research Resource [6, 7]. RRIs were extracted from ECG signals using a delineation algorithm [8]. Sleep stage annotations in $30\,\mathrm{s}$ epochs provided in SHHS1 were aligned with the extracted RRIs. A rolling median filter (51-beat window) was applied to the RRI series. Beats outside $(0.85n < \mathrm{RRI} < 1.15n)$, where n is the local median, were removed. Subjects with over 20% of RRIs excluded were discarded from analysis. The resulting dataset comprised RRI data from 4158 subjects. We further excluded subjects with sleep apnea to avoid confounding effects in sleep stage analysis. The final cohort comprised 2052 subjects, summarized in Table 1.

Majority-vote smoothing [9] was applied to $5\,\mathrm{min}$ segments to reduce fluctuations from $30\,\mathrm{s}$ epoch scor-

Table 1: Demographic characteristics of the study subjects. Age and body mass index (BMI) are presented as mean ± standard deviation.

Variable	Value
Subjects N (male/female)	2052 (684/1368)
Age (years)	60.7 ± 11.0
BMI (kg/m²)	26.9 ± 4.4

ing. Sleep stages were simplified into three macrostates—WAKE, REM, and NREM in order to reduce inter-rater variability, balance class distributions, and focus on physiologically distinct states relevant to autonomic and cardiovascular dynamics [10].

The dataset of 2052 subjects was randomly partitioned into a training set (80%) and an independent test set (20%). During model development, the training set was further split into training and validation subsets (80/20) to enable parameter optimization, while the test set remained untouched until final training and evaluation.

3. Theory and methods

DFA is a common nonlinear method for HRV analysis, and here we focus on its dynamical extension, DDFA [5]. This method enables the assessment of scaling exponents $\alpha(t, s)$ as functions of both time and scale, providing a detailed view of temporal dynamics. It has been successfully applied to sleep stage identification in earlier work [11].

Second-order DDFA was applied to RRIs across 38 temporal scales ranging from 6 to 499 heartbeats over the entire recording. The resulting scaling exponents $\alpha(t, s)$ were segmented into 30 s epochs aligned with the sleep stage annotations. For each epoch, $\alpha(t, s)$ values were estimated across all scales, and their means and standard deviations were computed to characterize epoch-level dynamics. This process yielded 76 features in total, corresponding to the mean and standard deviation of $\alpha(t, s)$ at each of the 38 scales.

Fig. 1 illustrates the overnight DDFA scaling exponent for a single subject, with blue regions indicating anti-correlations and red regions indicating positive correlations between heartbeats. While general trends are evident across sleep stages, particularly the blue regions at high scales during NREM sleep, considerable variability exists within each stage.

cHRV features were extracted from the same 30 s epochs. Time-domain features included the mean RRI and RMSSD, while frequency-domain features consisted of LF and HF and their ratio LF/HF. In addition, we also calculated the Poincaré plot indices SD1 and SD2. These metrics collectively formed the cHRV feature space [4].

Time-domain and Poincaré HRV features were computed directly within 30 s sleep-stage epochs. Since frequency-domain metrics require longer data, LF/HF was estimated from 5 min windows and assigned to the central 30 s. Similarly, DDFA features were derived from extended RR sequences to ensure reliable estimation at larger

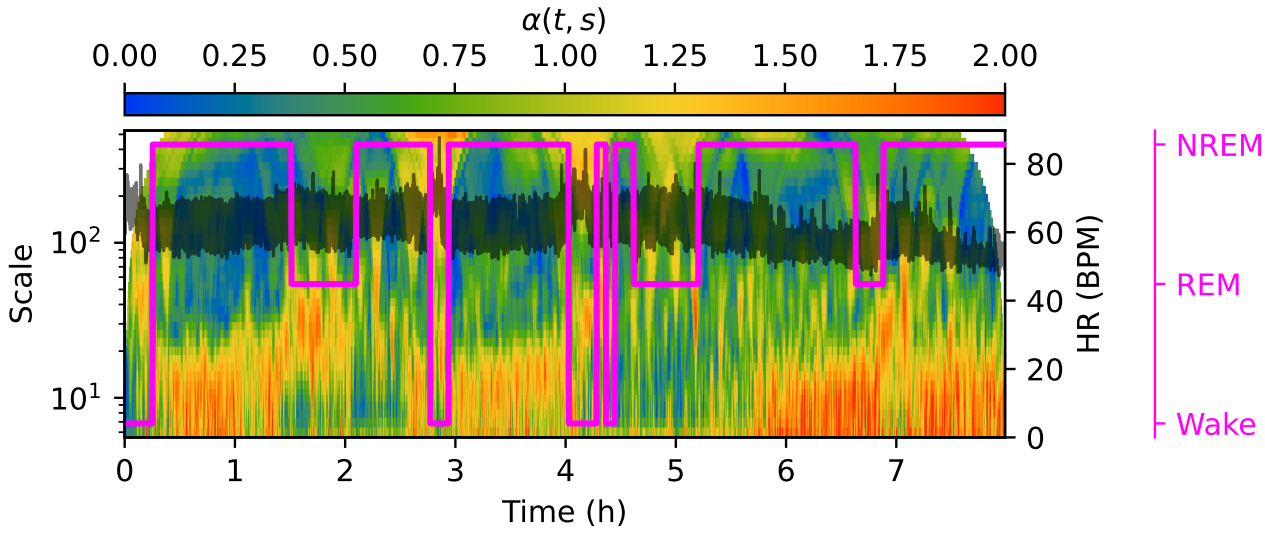

Figure 1: Overnight DDFA landscape for a single subject. The black curve depicts heart rate, and the pink line shows majority-vote smoothed sleep stages. The plot presents time on the x-axis and scale on the y-axis, with colors representing DDFA scaling exponents $\alpha(t, s)$. Blue regions in higher scales indicating anti-correlations align closely with NREM sleep periods.

scales. This approach enhances feature validity but may smooth rapid stage transitions, as both LF/HF and large-scale DDFA reflect information across multiple epochs.

Both DDFA and cHRV metrics were used independently as input features for three machine learning classifiers: multinomial logistic regression (LogReg, scikit-learn [12]), eXtreme Gradient Boosting (XGBoost [13]), and a neural network (NN, PyTorch [14]). Sleep stages were treated as class labels, and the classifiers' performance was compared. Class imbalance was handled by applying training-set weights.

The logistic regression baseline was implemented as a multinomial classifier using a softmax output layer. Input features were standardized to zero mean and unit variance. The model was trained with the saga solver, using a maximum of 500 iterations and an inverse regularization strength $C = 1.0$.

The XGBoost classifier was tuned via 3-fold cross-validation using grid search over tree depth, learning rate, number of estimators, and sampling rates. Optimal hyperparameters differed between feature spaces: for DDFA, 800 estimators, maximum depth 8, learning rate 0.1, subsample 0.8; for cHRV, 800 estimators, maximum depth 4, learning rate 0.05, subsample 0.8. Both models used all available features when building each tree.

The NN consisted of two hidden fully-connected layers with 32 neurons, ReLU activation and a dropout of $p = 0.1$. The optimizer was set to Adam [15] with an initial learning rate of 0.001, along with an exponential learning rate decay with $\gamma = 0.9$.

4. Results

Figure 2 presents the confusion matrices for all three classifiers: logistic regression, XGBoost, and the neural network, using both feature sets, cHRV and DDFA. For logistic regression, wake detection performance is similar, 0.53 for cHRV and 0.54 for DDFA. DDFA improves REM detection from 0.33 to 0.55. Likewise, DDFA NREM detection is improved from 0.48 to 0.68 compared to cHRV as seen in Figs. 2(a,b).

Figures 2(c,d) show the performance of the XGBoost algorithm. DDFA improved wake detection from 0.40 to 0.52 and REM detection from 0.49 to 0.64. NREM classification also improved, increasing from 0.51 to 0.73 compared to cHRV.

For the neural network, cHRV performed better in wake detection, achieving a score of 0.59 compared to 0.48 with DDFA. In contrast, DDFA improved REM detection markedly from 0.29 to 0.66, and REM classification increased from 0.49 to 0.73 as illustrated in Figs. 2(e,f).

Different classifier architectures had minimal impact on overall performance, with balanced accuracy varying by only two percentage points for cHRV and four percentage points for DDFA, although some models performed better on specific sleep stages. This suggests that classification performance is primarily limited by the feature space rather than the choice of architecture.

Similar trends were observed in preliminary experiments with convolutional neural network architectures, where increasing model size did not substantially improve performance. Furthermore, additional dataset augmentation, excluding approximately 400 subjects with a history of cardiovascular diseases, did not affect model performance.

Figure 2: Conventional HRV (cHRV) parameters (left) and DDFA feature space (right) classification results for (a,b) logistic regression, (c,d) XGBoost, and (e,f) neural network.

5. Conclusion

Across all tested classifiers and feature sets, DDFA consistently outperformed cHRV, indicating that the feature representation, rather than model complexity, was the key

determinant of classification performance. This suggests that DDFA provides a more informative and discriminative characterization of heartbeat dynamics for sleep stage classification.

By explicitly capturing correlations between consecutive heartbeats, DDFA produces features that are both explainable and intuitive. These features can be easily visualized, which makes them particularly well suited for use in interpretable machine learning and explainable AI frameworks.

Because HRV can be measured non-invasively using photoplethysmography [11], DDFA offers a practical alternative to PSG. Its adaptability to wearable devices such as smart rings, watches, and belts highlights its potential for real-world sleep monitoring and wellness applications.

The present results highlight the importance of the selected feature space in determining classification performance. In future work, the feature space could be refined by utilizing the raw values of $\alpha(t, s)$ rather than summarizing them with means and standard deviations. The present study also motivates further research on evaluating DDFA performance in sleep disorders, particularly across different levels of sleep apnea severity. Such analyses will help determine how apnea burden influences classification accuracy and the broader clinical applicability of DDFA-based approaches.

Acknowledgments

This work was supported by the Finnish Doctoral Program Network in Artificial Intelligence AI-DOC (VN/3137/2024-OKM-6), KAUTE foundation (20240420), and Finnish Foundation for Technology Promotion (10146). We are grateful to Matti Molkkari for supplying us with the computer program used to calculate DDFA.

The Sleep Heart Health Study (SHHS) was supported by National Heart, Lung, and Blood Institute cooperative agreements U01HL53916 (University of California, Davis), U01HL53931 (New York University), U01HL53934 (University of Minnesota), U01HL53937 and U01HL64360 (Johns Hopkins University), U01HL53938 (University of Arizona), U01HL53940 (University of Washington), U01HL53941 (Boston University), and U01HL63463 (Case Western Reserve University). The National Sleep Research Resource was supported by the National Heart, Lung, and Blood Institute (R24 HL114473, 75N92019R002).

References

[1] Memar P, Faradji F. A novel multi-class EEG-based sleep stage classification system. IEEE Transactions on Neural Systems and Rehabilitation Engineering Jan 2018; 26(1):84–95.

[2] Labarca G, et al. CPAP in patients with obstructive sleep apnea and type 2 diabetes mellitus: systematic review and meta-analysis. Clinical Respiratory Journal Aug 2018; 12(8):2361–2368.

[3] Rundo JV, Downey R. Chapter 25 - Polysomnography. In Levin KH, Chauvel P (eds.), Clinical Neurophysiology: Basis and Technical Aspects, volume 160 of Handbook of Clinical Neurology. Elsevier, 2019; 381–392.

[4] Shaffer F, Ginsberg JP. An overview of heart rate variability metrics and norms. Frontiers in Public Health 2017;5:258–258.

[5] Molkkari M, et al. Dynamical heart beat correlations during running. Scientific Reports 2020;10(1):13627.

[6] Zhang GQ, et al. The National Sleep Research Resource: towards a sleep data commons. Journal of the American Medical Informatics Association May 2018; 25(10):1351–1358.

[7] Quan SF, et al. The Sleep Heart Health Study: design, rationale, and methods. Sleep 1997;20(12):1077–1085.

[8] Niemi T. Computational optimization of wave detection algorithms for the electrocardiogram. Master's thesis, Tampere University, 2024.

[9] Guillot A, et al. Dreem open datasets: multi-scored sleep datasets to compare human and automated sleep staging. IEEE Transactions on Neural Systems and Rehabilitation Engineering September 2020;28(9):1955–1965.

[10] Almutairi H, et al. Machine-learning-based-approaches for sleep stage classification utilising a combination of physiological signals: a systematic review. Applied Sciences December 2023;13(24):13280.

[11] Molkkari M, et al. Non-linear heart rate variability measures in sleep stage analysis with photoplethysmography. Computing in Cardiology 2019;46.

[12] Pedregosa F, et al. Scikit-learn: machine learning in Python. Journal of Machine Learning Research 2011; 12:2825–2830.

[13] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016; 785–794.

[14] Paszke A, et al. PyTorch: an imperative style, high-performance deep learning library, 2019.

[15] Kingma DP, Ba J. Adam: a method for stochastic optimization 2014;.

Address for correspondence:

Topi Niemi
Computational Physics Laboratory, Tampere University
P.O. Box 600, FI-33014 Tampere, Finland
topi.niemi@tuni.fi