Explainable Sleep Stage Classification Based on ECG-Derived HRV Features: A Comparative Study of Machine Learning and Deep Learning Approaches

Junjia Long *

School of Health Sciences & Biomedical Engineering, Hebei University of Technology, Tianjin, China

* Corresponding Author Email: looong.rw@gmail.com

Abstract. Automatic sleep staging from wearable-compatible signals still runs into a basic trade-off: accessibility versus diagnostic detail. This study asks whether heart rate variability (HRV) from single-lead Electrocardiogram (ECG) can support reliable four-stage sleep classification (Wake, Light Sleep, Deep Sleep, REM), and whether temporal modeling plus explainable Artificial Intelligence (AI) can push both accuracy and clinical trust further. This study evaluated two models on 129,739 HRV windows from 148 subjects in the Haaglanden Medisch Centrum (HMC) sleep staging database, using the same subject-independent split for both. Random Forest (RF) worked from independent HRV vectors. Bidirectional Long Short-Term Memory (Bi-LSTM) with multi-head attention (Bi-LSTM+Attention) worked from sequential HRV representations instead. Bi-LSTM+Attention won on every metric: accuracy 0.453 versus 0.409, macro F1 0.445 versus 0.382, kappa 0.257 versus 0.164. The biggest gains showed up in wake and REM, the two most underrepresented classes in the data. SHapley Additive exPlanations (SHAP) analysis found that both models leaned heavily on autonomic-related features, particularly Detrended Fluctuation Analysis (DFA) α1, Low Frequency / High Frequency (LF/HF) power, and CV_RR, though Bi-LSTM gave more weight to short-term parasympathetic markers like Root Mean Square of Successive Differences (RMSSD) and SD1, which fits with its access to temporal context that RF simply doesn't have. Overall, ECG-derived HRV looks genuinely useful for automatic sleep staging when paired with sequential deep learning and model interpretation. It hasn't caught up to full polysomnography yet, but the framework shows real promise for practical use in wearable and home-based monitoring.

Keywords: Sleep stage classification, Electrocardiogram, Heart rate variability, Bi-LSTM, SHAP.

1. Introduction

Sleep stage classification is central to assessing sleep quality and tracking the physiological changes that unfold across a night. In clinical practice, polysomnography (PSG) is still the gold standard, mainly because it captures such comprehensive physiological data. But it needs specialized equipment and a controlled recording environment, which rules it out for long-term, home-based, or wearable monitoring[1]. Some recent wearable approaches have tried to work around this using continuous physiological sensing instead[2]. That's part of why interest has grown in automatic classification methods built on more accessible signals, ones that could ease the burden of traditional PSG-based assessment.

Electrocardiography (ECG) is a natural fit here. It's non-invasive and works well with wearable hardware. Heart rate variability (HRV), calculated from consecutive RR intervals in the ECG, reflects autonomic nervous system activity, and it shifts in recognizable ways across different sleep stages [3]. That physiological grounding is what makes ECG-derived HRV a reasonable basis for automatic sleep staging in the first place.

Handcrafted HRV features, spanning time-domain, frequency-domain, and nonlinear measures, have worked reasonably well with conventional machine learning. Random Forest in particular gets used a lot here, largely because it handles physiological feature sets robustly[4]. The catch is that these models usually score each sleep epoch on its own, without accounting for how adjacent stages relate to each other[4]. That independence assumption can make it harder to catch the gradual transitions that actually characterize sleep.

LSTM-based architectures were built to get around exactly this kind of limitation, learning dependencies across sequential data rather than treating each point in isolation, which suits physiological time series well[5]. End-to-end deep models applied straight to single-lead ECG have shown similar gains from capturing temporal structure[6]. Since HRV itself evolves continuously through the night, feeding a model sequential HRV representations should, in principle, help it pick up on transitions between adjacent epochs. Even so, direct comparisons between traditional machine learning and deep learning on ECG-derived HRV remain fairly limited, and how much temporal modeling actually buys you in this setting is still an open question.

Interpretability is a separate problem, and a real one in biomedical AI. Deep learning models often perform well while staying opaque about which features actually drove a given prediction[7]. Attention-based architectures on multimodal physiological signals have started to address this, aiming for predictive performance and some degree of interpretability at once[8]. Similar attention-based combinations of ECG and EEG have reportedly improved staging accuracy too[9]. This matters more in clinical sleep analysis than it might elsewhere, since predictions clinicians can't explain tend to erode trust and limit real-world use. Knowing which HRV features actually drive a classification helps connect the model's behavior back to physiology clinicians already understand.

This study works through these issues using a comparative framework built on the Haaglanden Medisch Centrum (HMC) sleep staging database, a clinical PSG dataset of 151 whole-night recordings with ECG signals annotated under AASM guidelines[10]. It makes three contributions. First, it builds an HRV feature-based pipeline to test whether ECG-derived information is viable for automatic sleep staging in a clinical setting. Second, it compares Random Forest on independent HRV representations against Bi-LSTM with attention on sequential HRV representations, to see what temporal modeling actually buys for ECG-based staging. Third, it applies SHAP analysis to identify which HRV features matter most and to ground model decisions in known autonomic physiology. Together, these pieces give a structured look at how traditional machine learning stacks up against deep learning here, while also making ECG-based sleep analysis more interpretable along the way.

2. Materials and Methods

The overall workflow of this study, from raw ECG acquisition to SHAP-based model interpretation, is illustrated in Fig. 1.

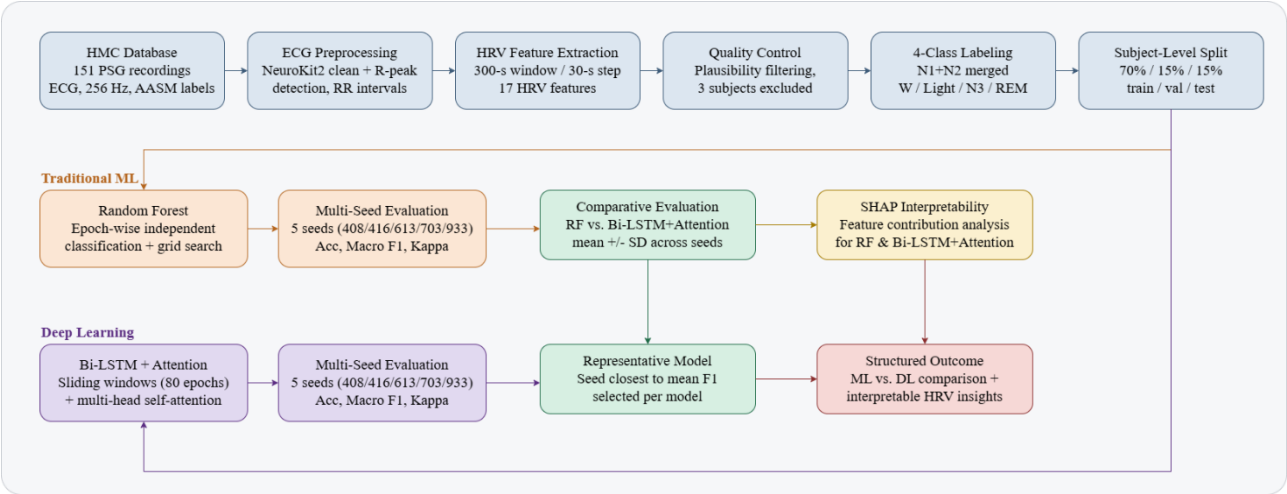


Fig. 1 Overall framework of ECG-based sleep stage classification (Picture credit: Original)

2.1. Dataset Description

This study uses the Haaglanden Medisch Centrum (HMC) sleep staging database, a publicly available polysomnographic dataset hosted on PhysioNet. The dataset contains 151 whole-night PSG recordings collected in 2018 at the HMC sleep center in the Netherlands. Each recording includes

multiple physiological signals, including four EEG channels (F4-M1, C4-M1, O2-M1, C3-M2), two EOG channels, one chin EMG channel, and one single-lead ECG channel, while the ECG signal used in this study was analyzed at its original sampling frequency of 256 Hz.

Sleep stages were annotated by trained sleep technicians according to the AASM 2.4 guidelines[11]. In this study, the original five-stage sleep labels were merged into four classes by combining N1 and N2 as light sleep: wakefulness (W), light sleep (N1+N2), deep sleep (N3), and REM sleep (R).

The original sleep scoring was performed using 30-second epochs, resulting in 137,243 labeled epochs across all subjects. In this study, these 30-second annotations were retained as the temporal resolution of sleep stage classification, while ECG-derived HRV features were calculated using a longer sliding-window strategy to improve physiological representation.

Sleep stages were originally scored in 30-second epochs, giving us 137,243 labeled epochs across the full cohort. The 30-second resolution was retained for the classification task itself, but calculated the HRV features using a longer sliding window to get a more physiologically meaningful representation.

The database started with 151 subjects. One subject, SN098, dropped out during HRV extraction because ECG-derived feature generation failed for that recording. Physiological plausibility filtering removed a second, SN130, since none of its windows fell within acceptable ranges. A third, SN036, came out during subject-level quality control: too few valid temporal samples remained for reliable sequence modeling. That left 148 subjects and 129,739 valid HRV windows for model development. N1 and N2 were also merged into a single light sleep category, since ECG-derived HRV features can't reliably tell the two apart. The subsequent experiments were therefore conducted on four sleep stages, namely wakefulness (W), light sleep (N1+N2), deep sleep (N3), and REM sleep (R). After this adjustment, the dataset contained 16.7% W, 48.1% light sleep, 19.7% N3, and 15.6% R (Table 1).

Table 1. Stage distribution of the HMC dataset after preprocessing

Stage	Epochs
W	21,114
N1+N2	62,467
N3	25,794
R	20,415

2.2. ECG Signal Processing

We pulled the single-lead ECG from each PSG recording, sampled at 256 Hz, and preprocessed it with NeuroKit2[12]. The raw signal was cleaned with the ecg_clean function, then R-peaks were detected using ecg_peaks. Full waveform delineation was not required here, since HRV analysis only requires R-peak locations. RR intervals came from the time between consecutive R-peaks, converted to milliseconds.

To get stable HRV estimates, a centered sliding window was applied. For each 30-second sleep epoch, a 300-second ECG segment was extracted, centered on that epoch's midpoint and computed HRV from it, then assigned the resulting features back to that epoch's stage label. This gives frequency-domain and nonlinear HRV measures enough cardiac cycles to work with, without sacrificing the 30-second resolution the staging task needs.

Windows with too few valid RR intervals or unreliable HRV estimates were excluded, and filtered further using physiologically plausible ranges for each HRV parameter to catch artifacts from R-peak detection errors.

2.3. HRV Feature Extraction

Each 5-minute ECG window yielded 17 HRV features spanning time-domain, frequency-domain, and nonlinear measures.

Seven time-domain features captured RR interval variability: mean RR interval, SDNN, RMSSD, NN50, pNN50, mean heart rate, and CV_RR. Each reflects a different angle on cardiac autonomic regulation. SDNN tends to track overall variability, while RMSSD and pNN50 lean more toward short-term parasympathetic activity[13]. and all of these are standard tools in sleep-related autonomic analysis.

Frequency-domain features came from the spectral profile of the RR interval series. Using 5-minute windows instead of 30-second segments gave us enough data for reliable frequency decomposition. This set included VLF, LF, and HF power, the LF/HF ratio, and normalized LF and HF power, all of which have been linked to shifts in autonomic regulation across different physiological states during sleep.

For nonlinear dynamics, Poincaré plot analysis was applied to get SD1 and SD2, capturing short-term and long-term RR variability respectively[14], plus their ratio to gauge how much short-term variability contributes relative to the rest. The DFA α1 scaling exponent was also computed to characterize short-range fractal correlations in RR interval dynamics[15].

All 17 features were combined into a single feature vector per temporal sample. These vectors then fed two separate classification approaches: RF, working from independent HRV vectors, and Bi-LSTM, built to model temporal dependencies across sequential HRV representations.

2.4. Sleep Stage Classification Framework

2.4.1. Random Forest Based on HRV Features

Random Forest (RF) was implemented as the baseline machine learning classifier[16] using the extracted 17-dimensional HRV feature vectors. Each temporal sample was treated as an independent input without considering the information from neighboring epochs. This design allowed the evaluation of classification performance based solely on instantaneous autonomic characteristics represented by HRV features.

Before model training, HRV features were standardized using z-score normalization based on the training set statistics. The RF classifier consisted of multiple decision trees trained using bootstrap sampling and random feature selection. Hyperparameter optimization was performed on the validation set, including the number of trees, maximum tree depth, and minimum samples per leaf.

The optimal configuration was determined as 300 estimators, unlimited maximum depth, and a minimum leaf size of 3. Class imbalance among sleep stages was addressed using balanced class weighting during training. The final model was evaluated on the independent test set using accuracy, macro-averaged F1 score, and Cohen’s kappa coefficient.

2.4.2. Bi-LSTM Based on Sequential HRV Features

To incorporate temporal dependencies between consecutive sleep epochs, a bidirectional long short-term memory (Bi-LSTM) network with a multi-head self-attention layer was developed[5]. Unlike RF, which scores each 30-second epoch on its own, this architecture works with sequences of HRV vectors, so a given epoch's prediction can draw on both what came before it and what comes after.

Each subject's HRV sequence was divided into fixed windows of 80 consecutive epochs, 40 minutes, sliding forward 40 epochs at a time. Where a recording ended before filling out a full window, the remainder was zero-padded, with a binary mask applied to exclude those padded positions from the loss and evaluation calculations.

The network itself is a two-layer bidirectional LSTM followed by multi-head self-attention. The Bi-LSTM reads the HRV sequence forward and backward to build temporal representations, then the attention layer weighs each time step differently before the model makes its final prediction[17]. In this way, the model considers the entire input sequence instead of treating each epoch independently. A fully connected output layer produced epoch-level class predictions for each time step in the sequence.

Model training used a class-weighted cross-entropy loss, with weights computed from the training set's label distribution to mitigate the impact of class imbalance, particularly the underrepresentation

of the wakefulness and REM classes relative to light sleep. Model hyperparameters, including the hidden dimension, LSTM depth, number of attention heads, dropout rate, and learning rate, were determined through validation experiments. Training was terminated early when the validation Macro F1 score showed no improvement for four consecutive epochs, and the model achieving the highest validation performance was retained for subsequent evaluation.

Because the performance of deep neural networks can vary with random initialization, the final Bi-LSTM model was trained five times using the same random seeds as the Random Forest experiments (408, 416, 613, 703, and 933) , with the average and standard deviation reported across these five runs. For SHAP analysis, seed 933 was picked as the representative model since its performance sat closest to that average[18]. Training and validation curves for all five seeds appear in Fig. 2, and convergence looked consistent across the board.

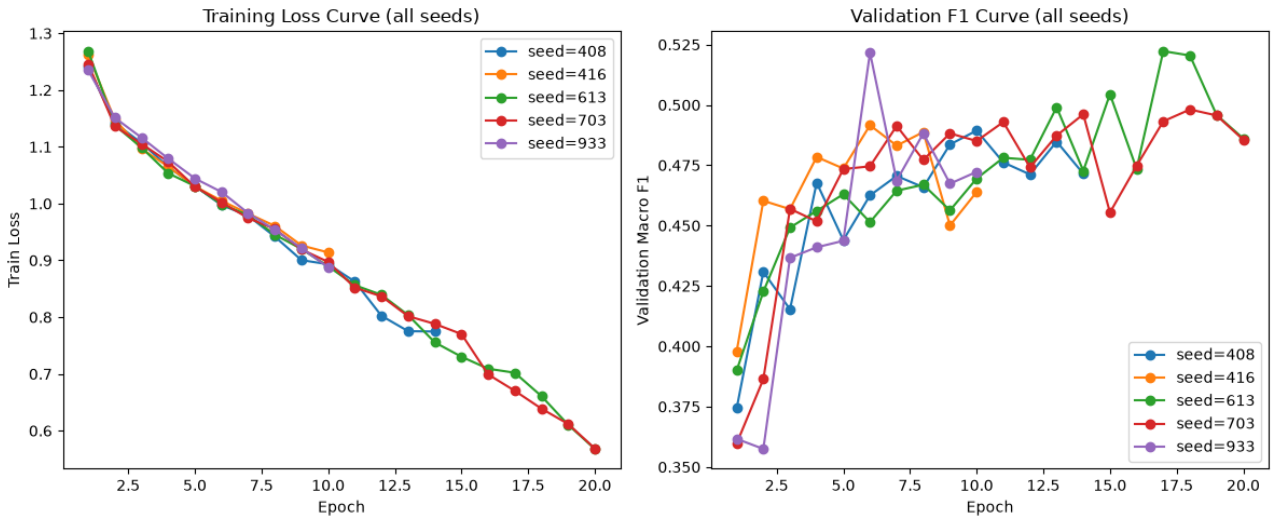


Fig. 2 Training and validation performance across five random seeds (Picture credit: Original)

2.5. Performance Evaluation

The performance of the proposed sleep stage classification models was evaluated using three commonly used metrics: accuracy, macro-averaged F1 score, and Cohen's kappa coefficient.

Accuracy was calculated as the percentage of correctly classified epochs among all valid test samples. Although accuracy can be used as an overall measure of classification performance, it will be affected by category imbalance. In sleep staging, light sleep usually accounts for the largest proportion of the data.

To evaluate all sleep stages more fairly, Macro F1 was used as the primary evaluation metric. Macro F1 combines precision and recall while giving equal weight to each sleep stage, regardless of the number of samples. This metric is therefore more suitable for evaluating minority classes such as wakefulness and REM sleep.

Cohen's kappa coefficient was also calculated to measure the agreement between model predictions and manual annotations after accounting for chance agreement. A higher kappa value indicates stronger consistency between the predicted and manually scored sleep stages.

For Random Forest, performance was evaluated using the final model trained on the training set and optimized through validation-based hyperparameter selection. For the Bi-LSTM with attention model, considering the sensitivity of deep learning models to initialization, five independent training runs with different random seeds were performed. The final performance was reported as the mean ± standard deviation across the five runs. In addition, a representative model with performance closest to the average across seeds was selected for subsequent confusion matrix analysis and SHAP-based interpretability analysis.

3. Results

3.1. Classification Performance

The performance of the proposed ECG-based sleep stage classification framework was evaluated on the independent test set using accuracy, macro-averaged F1 score, and Cohen's kappa coefficient. Both models were trained and tested under the same subject-independent split.

The RF classifier reached an average accuracy of 0.4087 ± 0.0014, macro F1 of 0.3823 ± 0.0019, and kappa of 0.1637 ± 0.0025 across five seeds. The variation between runs was small. RF, after all, has no randomness in initialization the way a neural network does, so this stability was expected.

Bi-LSTM+Attention did better across the board, averaging 0.4529 ± 0.0379 accuracy, 0.4453 ± 0.0329 macro F1, and 0.2570 ± 0.0322 kappa. Accuracy climbed 4.4 points over RF. Macro F1 climbed further, by 6.3 points. That gap between the two gains matters: it points to the Bi-LSTM picking up ground mainly on the minority stages, not just riding the dominant light sleep class to a better score.

Among the five repeated experiments, the best performance was obtained with seed 703, achieving an accuracy of 0.4900, macro F1 score of 0.4788, and Cohen's kappa coefficient of 0.2983. These results indicate moderate agreement between the predicted labels and the manual annotations. The representative model selected for further interpretation analysis was seed 933, whose performance was close to the overall average, achieving an accuracy of 0.4668, macro F1 score of 0.4495, and Cohen's kappa coefficient of 0.2574.

The better performance of the Bi-LSTM suggests that temporal information is helpful for ECG-based sleep staging. Sleep stages change gradually during the night. By using consecutive epochs together, the Bi-LSTM can capture these transitions, whereas RF classifies each epoch independently.

3.2. Confusion Matrix Analysis

Fig. 3 shows the confusion matrix for the representative Bi-LSTM+Attention model (seed 933). Performance varied a lot by stage.

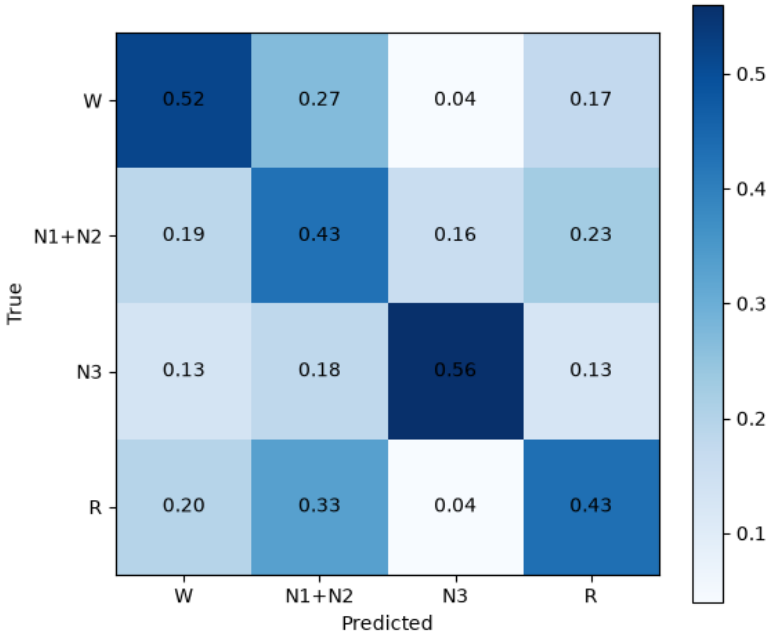


Fig. 3 Confusion matrix of the Bi-LSTM with attention model on the test set (seed=933) (Picture credit: Original)

N3 was the easiest stage for the model to identify, with 56.0% recall and an F1 of 0.55. That's probably down to how stable deep sleep is physiologically: heart rate drops, parasympathetic activity takes over, and the HRV signature that results stays fairly consistent epoch to epoch, giving the model less noise to work through.

The light sleep category (N1+N2) achieved the highest precision (64%) but relatively lower recall (42.8%). As the largest class in the dataset by far, it also absorbed a lot of misclassified epochs from wake, N3, and REM. That's not surprising. N1 and N2 are transitional states, physiologically heterogeneous, and their autonomic signatures overlap with whatever stage they're bordering.

Wake recall landed at 51.6%, so roughly half of true wake epochs were caught. Most of the errors went to light sleep, which lines up with how similar quiet wakefulness and light sleep already look in terms of heart rate regulation.

REM was the toughest stage to pin down, with recall at just 43.3% and most errors again landing in light sleep. REM does have a recognizable autonomic fingerprint, rapid alternation between sympathetic and parasympathetic dominance, but that fingerprint isn't sharp enough on ECG alone to fully separate REM from its neighbors.

Overall, the confusion matrix shows that the Bi-LSTM recognized sleep-stage patterns better than the RF model, but considerable overlap remains between adjacent physiological states, particularly between light sleep and other stages.

3.3. Comparison Between RF and Bi-LSTM

A comparison between the conventional machine learning approach and the sequential deep learning approach is summarized in Table 2.

Table 2. Performance comparison of RF and Bi-LSTM models

Model	Accuracy	Macro F1	Kappa
RF	0.4087 ± 0.0014	0.3823 ± 0.0019	0.1637 ± 0.0025
Bi-LSTM + Attention	0.4529 ± 0.0379	0.4453 ± 0.0329	0.2570 ± 0.0322

Bi-LSTM+Attention beat RF on every metric. Most of that edge traces back to its ability to model temporal dependencies across consecutive HRV windows. RF looks at each 30-second epoch in isolation; the sequential model instead has access to a roughly 40-minute window of surrounding context, enough to pick up on gradual stage transitions rather than treating each moment as unrelated to the last.

However, the gain was modest, not dramatic. Kappa values in this range are broadly in line with what's been reported for other HRV or PPG-based staging work on comparable clinical cohorts[19]. Temporal modeling clearly helps, but HRV on its own may just not carry enough information to cleanly separate every stage. The persistent confusion around N1+N2 in particular suggests that closing this gap further would likely need additional modalities such as EEG or EOG, not just better temporal models.

The RF model achieved relatively stable but limited performance, demonstrating that handcrafted HRV features contain meaningful physiological information for sleep classification. Meanwhile, Bi-LSTM builds on that by learning temporal structure the handcrafted features alone can't express. Together, the two results make a reasonable case for temporal deep learning as a complement to, rather than a wholesale replacement for, feature-based approaches in wearable ECG sleep monitoring.

4. Discussion

4.1. Interpretation of Important HRV Features Using SHAP

To make the framework's decisions easier to interpret, SHAP analysis was applied to both RF and Bi-LSTM.

For the RF model, DFA α1, CV_RR, and the LF/VLF/HF power features carried the most global importance (Fig. 4). DFA α1 was particularly useful for telling wakefulness and REM apart, which tracks with how autonomic regulatory complexity fades as sleep deepens. N3 showed the most regular, predictable heartbeat dynamics of any stage.

The frequency-domain features split fairly cleanly along parasympathetic and sympathetic lines. HF power and HF norm dominated for N3 and REM, both stages marked by strong vagal activity, while LF and VLF power mattered more for light sleep, where the sympathetic-parasympathetic balance is still shifting rather than settled. Time-domain features rounded out the picture. CV_RR and mean HR track the general slowing of heart rate as sleep deepens, and NN50 and pNN50 pick up on the shrinking of short-term interval fluctuations across NREM stages.

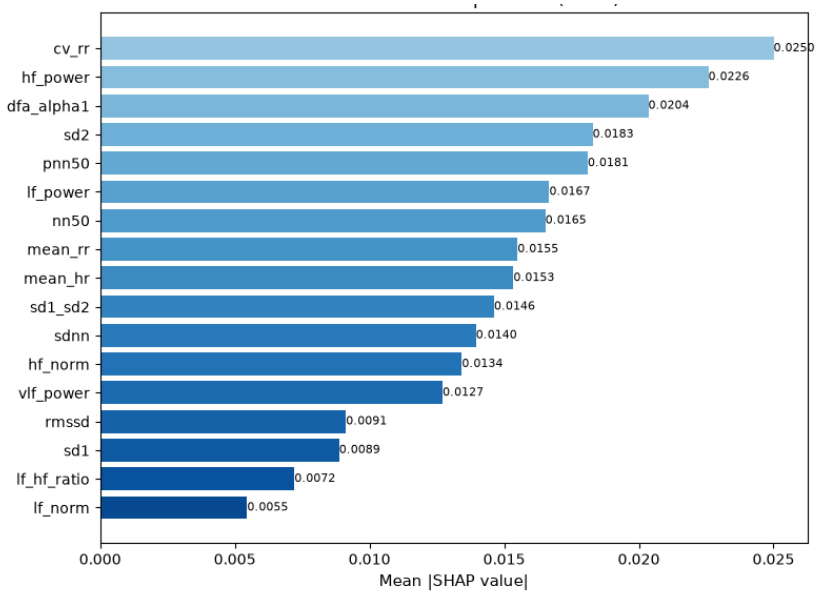


Fig. 4 Global SHAP feature importance of Random Forest (Picture credit: Original)

Bi-LSTM model (Fig. 5) leaned on many of the same features as RF, including DFA α1, LF power, VLF power, mean HR, and SD1/SD2, but the ranking shifted. RMSSD became one of the most influential features for wakefulness. HF norm and SD1 rose in importance for REM. This shift makes sense given that Bi-LSTM works with sequences rather than isolated samples: it can pick up on how HRV changes across consecutive epochs, so short-term parasympathetic markers like RMSSD and SD1, which only really mean something in context, become more useful than they are to a model scoring each epoch alone.

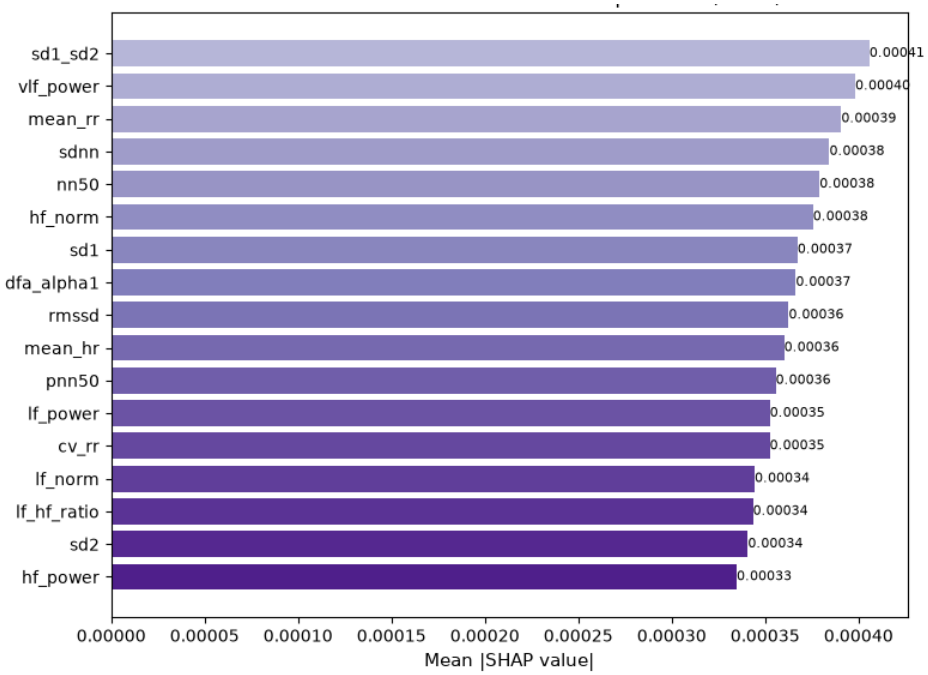


Fig. 5 Global SHAP feature importance of Bi-LSTM with Attention (Picture credit: Original)

Overall, both models leaned heavily on features tied to autonomic nervous system regulation. That's a reassuring sign. It means the classification results aren't just statistical artifacts, they line up with known sleep physiology.

4.2. Comparison Between Random Forest and Bi-LSTM

The two models used different strategies for sleep stage classification, which is reflected in both their performance and their SHAP importance profiles.

RF treats each epoch on its own, using handcrafted HRV descriptors from that window alone (per-stage SHAP importance in Fig. 6). This keeps the model simple and easy to interpret. The SHAP results spell out exactly which features drove each prediction, which is useful for clinical settings. But ignoring the relationship between adjacent epochs limits how well RF can model actual stage transitions, and that limitation shows up directly in its lower overall performance.

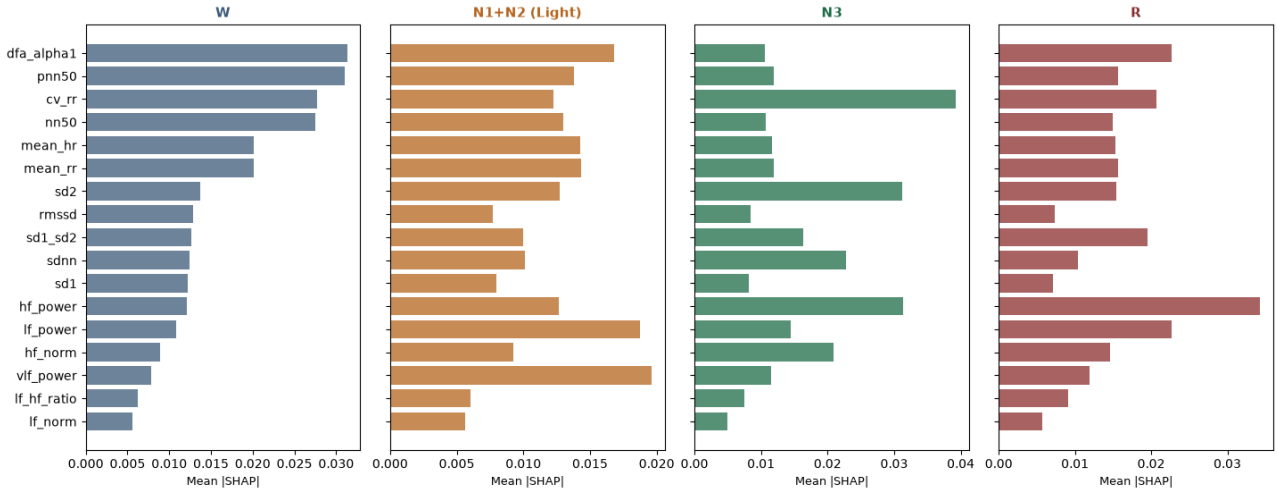


Fig. 6 Per-stage SHAP feature importance of Random Forest (Picture credit: Original)

Bi-LSTM's access to neighboring epochs is sharpened further by its multi-head attention layer (Fig. 7). Rather than weighting every time step the same, attention lets the model focus on the epochs where HRV is actually changing in a discriminative way. Around the transition into REM, for instance, autonomic activity shifts fairly abruptly, and the attention weights pick up on that shift.

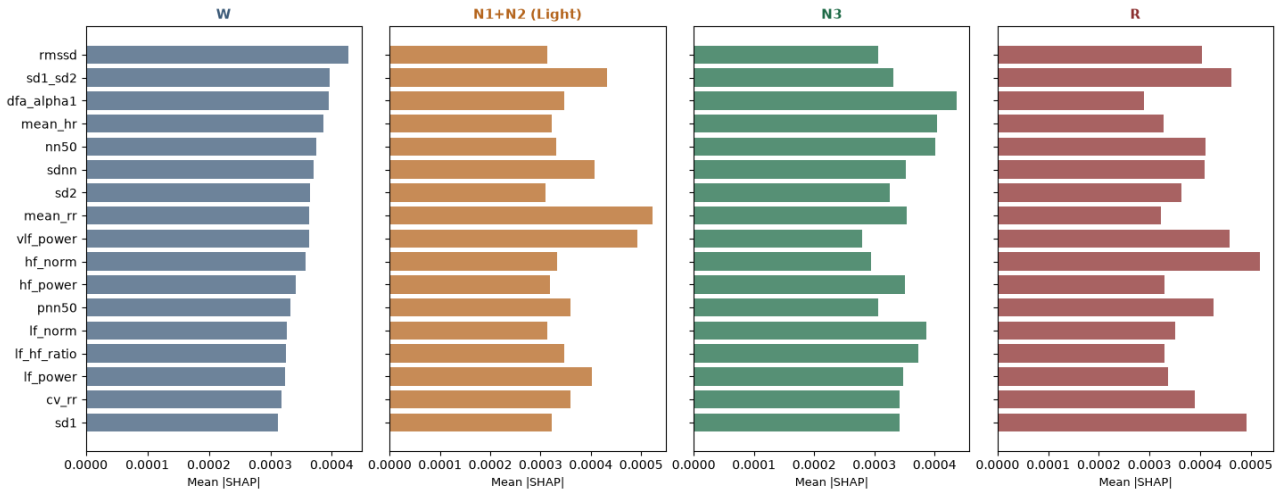


Fig. 7 Per-stage SHAP feature importance of Bi-LSTM with Attention (Picture credit: Original)

The SHAP analysis further illustrated this difference. While RF emphasized global statistical characteristics such as DFA α1 and CV_RR, the Bi-LSTM relied more heavily on features associated with short-term autonomic fluctuations, including RMSSD and SD1. These differences suggest that

the Bi-LSTM extracts dynamic temporal representations beyond the information available from isolated HRV measurements.

The SHAP comparison sharpens this contrast. RF leaned on stable, session-level features like DFA α1 and CV_RR. Bi-LSTM leaned more on short-term markers like RMSSD and SD1, features that only carry meaning relative to neighboring epochs. RF, without access to that context, simply couldn't make use of them the same way. That difference in what each model pays attention to, more than the raw performance gap, is the real interpretive payoff of comparing the two SHAP profiles.

4.3. Limitations and Future Work

Three aspects of the current study warrant careful interpretation.

The most basic one concerns the signal itself: only ECG-derived HRV, and HRV doesn't directly capture cortical activity, which is what clinical sleep scoring is actually built on. That gap likely explains why stages with similar autonomic signatures stayed hard to separate regardless of the model applied.

The dataset is also narrower than ideal. All 148 subjects came from a single clinical cohort, referred for PSG in the context of suspected sleep disorders. Whether the HRV patterns identified here generalize to healthy populations, or to signals recorded outside a clinical setting, is genuinely an open question. The 80-epoch, 40-minute sliding window was a practical compromise rather than an ideal choice. A longer window would probably capture full sleep cycles, which typically run about 90 minutes, more completely. But with only 2,215 training windows to work with, lengthening the window starts to destabilize training rather than improve it. The seed-to-seed variance in Bi-LSTM+Attention performance, macro F1 ranging from 0.39 to 0.47 across five runs, is a fairly direct symptom of that limited training data.

Finally, this work only looked at handcrafted HRV descriptors. Combining them with automatically learned ECG representations, or adding other physiological signals like respiration and actigraphy, seems like a reasonable next step for boosting performance without losing what makes ECG-based staging practical in the first place[20].

The results nonetheless indicate that ECG-derived HRV, paired with interpretable machine learning and temporal deep learning, can support meaningful automatic sleep staging without needing the full setup of a laboratory PSG.

5. Conclusion

This study compared a Random Forest classifier against a Bi-LSTM network with multi-head attention for ECG-derived HRV-based sleep stage classification, using the HMC clinical dataset. Bi-LSTM outperformed RF on every metric, gaining roughly 4 accuracy points and more than 6 macro F1 points, evidence that modeling temporal dependencies across epochs provides a real, measurable benefit, especially for underrepresented stages like wakefulness and REM. SHAP analysis added a second layer to this story. Both models relied heavily on autonomic-related features such as DFA α1, LF/HF power, and CV_RR, but Bi-LSTM placed more weight on short-term parasympathetic markers like RMSSD and SD1, reflecting its ability to exploit temporal fluctuations in cardiac regulation that a model scoring isolated epochs simply can't see.

These gains come with caveats. Section 4.3 covers them in more detail, but the short version is that absolute performance remains moderate, and light sleep continues to blur into adjacent stages in ways single-lead ECG alone can't fully resolve. Bringing in multimodal signals such as respiration and actigraphy, and testing on larger and more diverse cohorts, looks like the most promising path toward closing that gap while keeping the practical advantages that make wearable ECG monitoring worth pursuing in the first place.

References

[1] Mathunjwa BM, Kor RYJ, Ngarnkuekool W, et al. A comprehensive review of home sleep monitoring technologies: smartphone apps, smartwatches, and smart mattresses. Sensors, 2025, 25(6): 1771.

[2] Birrer V, Elgendi M, Lambercy O, et al. Evaluating reliability in wearable devices for sleep staging. npj Digital Medicine, 2024, 7: 74.

[3] Tobaldini E, Nobili L, Strada S, et al. Heart rate variability in normal and pathological sleep. Frontiers in Physiology, 2013, 4: 294.

[4] Chakraborty S, Goyal M, Goyal P, Mishra P. Artificial intelligence-based automated sleep staging using heart rate variability: assessment of performance and clinical prospects. National Medical Journal of India, 2026, 39(4): 224-230.

[5] Radha M, Fonseca P, Moreau A, et al. Sleep stage classification from heart-rate variability using long short-term memory neural networks. Scientific Reports, 2019, 9: 14149.

[6] Choi I, Sung W. Single-channel ECG-based sleep stage classification with end-to-end trainable deep neural networks. Proceedings of the 45th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2023: 1-4.

[7] Yue H, Chen Z, Guo W, et al. Research and application of deep learning-based sleep staging: data, modeling, validation, and clinical practice. Sleep Medicine Reviews, 2024, 74: 101897.

[8] Luo Y, Li J, He K, et al. A hierarchical attention-based method for sleep staging using movement and cardiopulmonary signals. IEEE Journal of Biomedical and Health Informatics, 2023, 27(3): 1354-1363.

[9] Lyu J, Shi W, Zhang C, et al. A novel sleep staging method based on EEG and ECG multimodal features combination. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2023, 31: 4073-4084.

[10] Alvarez-Estevez D, Rijsman RM. Inter-database validation of a deep learning approach for automatic sleep scoring. PLOS ONE, 2021, 16(8): e0256111.

[11] Iber C, Ancoli-Israel S, Chesson AL, et al. The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications. Westchester: American Academy of Sleep Medicine, 2007.

[12] Makowski D, Pham T, Lau ZJ, et al. NeuroKit2: A Python toolbox for neurophysiological signal processing. Behavior Research Methods, 2021, 53(4): 1689-1696.

[13] Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology. Heart rate variability: standards of measurement, physiological interpretation, and clinical use. Circulation, 1996, 93(5): 1043-1065.

[14] Milovanović B, Marković N, Petrović M, et al. Associations of Poincaré plot-derived parameters with heart rate variability and autonomic reflex testing in a real-world clinical population. Diagnostics, 2026, 16(7): 1016.

[15] Peng CK, Havlin S, Stanley HE, et al. Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. Chaos, 1995, 5(1): 82-87.

[16] Breiman L. Random forests. Machine Learning, 2001, 45(1): 5-32.

[17] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Advances in Neural Information Processing Systems, 2017, 30: 5998-6008.

[18] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 2017, 30: 4765-4774.

[19] Kotzen K, Charlton PH, Salabi S, et al. SleepPPG-Net: A deep learning algorithm for robust sleep staging from continuous photoplethysmography. IEEE Journal of Biomedical and Health Informatics, 2023, 27(2): 924-932.

[20] Krauss D, Richer R, Küderle A, et al. Incorporating respiratory signals for machine learning-based multimodal sleep stage classification: a large-scale benchmark study with actigraphy and heart rate variability. Sleep, 2025, 48(9): zsaf091.