POLITECNICO DI TORINO
Repository ISTITUZIONALE

Interpretable HRV-Based Detection of REM Sleep Behaviour Disorder via Explainable Boosting Machines

Original
Interpretable HRV-Based Detection of REM Sleep Behaviour Disorder via Explainable Boosting Machines / Digiacomo, F., Mosca, U., Ciampa, I., Rechichi, I., Cicolin, A., Olmo, G.. - (2026), pp. 1096-1101. (39th IEEE International Symposium on Computer-Based Medical Systems, CBMS 2026 Limassol (CY) June 3-5 2026) [10.1109/CBMS69103.2026.00223].

Availability:
This version is available at: 11583/3016270 since: 2026-10-05T07:37:33Z

Publisher:
Institute of Electrical and Electronics Engineers

Published
DOI:10.1109/CBMS69103.2026.00223

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright
IEEE postprint/Author's Accepted Manuscript

©2026 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collecting works, for resale or lists, or reuse of any copyrighted component of this work in other works.

(Article begins on next page)

11 November 2026

Interpretable HRV-Based Detection of REM Sleep Behaviour Disorder via Explainable Boosting Machines

Federico Digiacomo
Polytechnic of Turin
Turin, Italy
federico.digiacomo@polito.it

Umberto Mosca
Polytechnic of Turin
Turin, Italy
umberto.mosca@polito.it

Ilaria Ciampa
Polytechnic of Turin
Turin, Italy
ilaria.ciampa@polito.it

Irene Rechichi
Polytechnic of Turin
Turin, Italy
irene.rechichi@polito.it

Alessandro Cicolin
University of Turin
Turin, Italy
alessandro.cicolin@unito.it

Gabriella Olmo
Polytechnic of Turin
Turin, Italy
gabriella.olmo@polito.it

***Abstract*—Rapid Eye Movement (REM) sleep behaviour disorder (RBD) is an early biomarker of neurodegeneration but requires resource-intensive video-polysomnography (v-PSG) for diagnosis. This study presents an interpretable pipeline for automated detection of idiopathic RBD (iRBD) using heart rate variability (HRV) features extracted from ECG during sleep. HRV metrics are computed separately for each sleep stage to capture stage-specific autonomic signatures. Selected features are used to train an Explainable Boosting Machine (EBM). The approach is evaluated on an age-matched cohort of 67 subjects (33 healthy controls and 34 iRBD) using nested Leave-One-Subject-Out cross-validation. Across ten random seeds, the REM stage yielded the best performance (AUC = 0.82 $\pm$ 0.02, accuracy = 0.75 $\pm$ 0.04). Four HRV features are consistently selected across folds, indicating stable autonomic markers of iRBD and highlighting the potential of single-lead ECG as a scalable screening tool.**

***Index Terms*—REM sleep behaviour disorder, heart rate variability, explainable boosting machine, interpretable machine learning, polysomnography, leave-one-subject-out**

I. Introduction

Neurodegenerative synucleinopathies currently lack disease-modifying therapies [1]. Rapid Eye Movement (REM) sleep behaviour disorder (RBD), characterised by loss of normal muscle atonia during REM sleep, is one of the strongest prodromal markers of synucleinopathy: up to 80–90% of individuals with idiopathic RBD (iRBD) eventually convert within 10–15 years [2]–[4]. Diagnosis relies on attended video-polysomnography (v-PSG) [4], a resource-intensive procedure confined to specialised sleep centres, motivating the search for automated biomarkers extractable from PSG signals or wearable devices.

A. HRV and Sleep Stages as a Window on Neurodegeneration

A well-documented early feature of synucleinopathies is autonomic nervous system involvement, driven by Lewy body pathology spreading to sympathetic and parasympathetic ganglia and brainstem nuclei [5], [6]. Heart rate variability (HRV), derived from single-lead electrocardiogram (ECG), is a validated non-invasive measure of cardiac autonomic modulation [7]. Because distinct sleep stages exhibit substantially different autonomic profiles (parasympathetic dominance in N3, sympathetic surges in REM, intermediate states in N1+N2 [8]) computing HRV features separately for each stage can reveal signatures that whole-night analyses obscure. Since HRV also declines with ageing [9], the cohort in this study is age-matched between groups ($p = 0.135$, Mann–Whitney U test).

B. State of the Art and the Remaining Gap

Several studies have investigated machine learning approaches to RBD detection from PSG signals, ranging from deep learning on full-night multi-channel recordings to composite methods combining electromyography (EMG), electroencephalography (EEG), and cardiorespiratory features [10]. Rechichi et al. [11] proposed an automatic REM Sleep Without Atonia (RSWA) detection pipeline on the public CAP Sleep Database and on a subset of the TURIN database used here, reaching 87% accuracy and 93% sensitivity using spectral EMG features; that work focused on muscular atonia markers, whereas the present study examines the complementary perspective of cardiac autonomic modulation via single-lead ECG. HRV-based features have also been explored as markers of autonomic dysregulation [12], yet prior work typically analyses HRV without sleep-stage stratification or relies on opaque classifiers incompatible with regulatory transparency requirements.

C. Explainable AI as a Clinical Necessity

Black-box models compromise clinician trust, obscure failure mechanisms, and conflict with emerging regulatory frame-

works for medical AI [13], [14]. Post-hoc methods provide approximate interpretations but do not alter the fundamental opacity of the underlying model. Explainable Boosting Machines (EBMs) [15], [16], belonging to the family of generalised additive models (GAMs), are intrinsically interpretable: each feature's contribution is directly readable from the model with no approximation required, at accuracy competitive with gradient-boosted trees on tabular data [15].

D. Contributions

- A sleep-stage-stratified HRV pipeline applied to an age-matched, class-balanced polysomnographic cohort of 67 subjects divided between healthy controls (HC) and iRBD.
- A systematic feature selection with redundancy pruning to identify the most informative and non-redundant HRV metrics.
- An EBM classifier providing interpretable per-feature shape functions that reveal the autonomic signatures distinguishing iRBD from HC.

II. Methods

Figure 1 illustrates the overall workflow of the proposed HRV-based iRBD detection pipeline. Briefly, ECG signals are first extracted and preprocessed, followed by HRV feature extraction and aggregation. Subsequently, feature selection and hyperparameter optimisation are performed before training the final EBM model and interpreting its predictions through the model's native explainability.

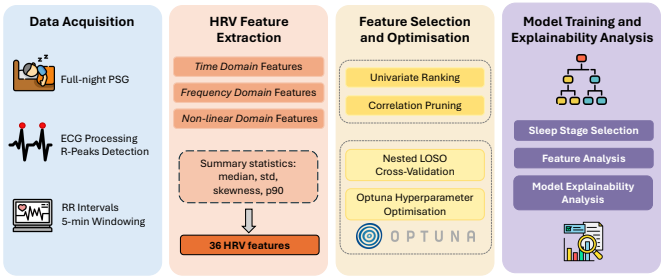


Fig. 1. Overview of the proposed HRV-based iRBD detection pipeline.

A. Dataset

In this study we considered a private database acquired at the Sleep Medicine Centre of Molinette Hospital, Turin, Italy (TURIN Database). The database comprises 67 full-night PSG recordings, including 33 HC and 34 iRBD. The study protocol was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of A.O.U. Città della Salute e della Scienza di Torino (approval number 00384/2020). Written informed consent was obtained from all participants prior to their inclusion in the study. Eligibility criteria included an age between 35 and 75 years, the absence of pharmacological treatments known to affect cardiac rhythm or alter ECG morphology or variability, and no diagnosis of neurodegenerative disease. Furthermore, iRBD patients included in the dataset did not present comorbid obstructive sleep apnoea syndrome (OSAS) or periodic limb movements (PLM). Table I summarises the demographic characteristics and sleep stage distribution of the study population.

For each PSG recording, sleep staging was performed using the corresponding hypnogram manually scored by experienced clinicians. Segments without an assigned sleep stage were excluded. For the following analyses, four sleep conditions were considered: REM sleep, N3, the combination of N1 and N2 stages, and the full recording after excluding wakefulness epochs (no-wake). This allowed HRV to be assessed across distinct sleep stages and over the total sleep period, enabling identification of the stage most informative for distinguishing iRBD from HC.

B. Signal Processing and RR Intervals Extraction

The ECG signals extracted from the PSG recordings were preprocessed using the following pipeline for HRV analysis. First, all recordings were resampled to 256 Hz to ensure a uniform temporal resolution across subjects, as the ECG signals were originally acquired using two different devices. A fourth-order Butterworth bandpass filter (5–30 Hz) was then applied to enhance the QRS complex while attenuating noise and low-frequency artefacts. R-peak detection was performed using the Python library NeuroKit2 [17]. To enhance signal reliability, the automatically detected peaks were further refined through dedicated correction procedures aimed at identifying and rectifying potential misdetections, thus maintaining the physiological integrity of the beat-to-beat sequence. RR intervals were computed as the time differences between consecutive R-peaks, generating the RR time series for subsequent analysis.

C. HRV Feature Extraction and Aggregation

HRV analysis was performed using a sliding-window approach with 5-minute windows and 30% overlap. Within each window, time-domain, frequency-domain, and non-linear HRV metrics were computed using the NeuroKit2 library [17]. Nine HRV features were selected based on their ability to quantify global autonomic failure and vagal tone. To prepare the input for the EBM, since the number of HRV windows varies across subjects, the per-window features were collapsed into a fixed-length subject-level vector by applying summary statistics across windows, resulting in 36 features per subject. The extracted features and the corresponding summary statistics are listed in Table II.

D. Feature Selection

To identify the most discriminative features while preserving the interpretability of the EBM's additive structure, a two-phase filter procedure is applied. Both phases operate exclusively on training data.

1) Univariate Ranking: Each aggregated HRV feature f is scored with a non-parametric Separability Index:

$$SI_f = \frac{\left|\tilde{x}_f^{(0)} - \tilde{x}_f^{(1)}\right|}{0.5\left(\mathrm{IQR}_f^{(0)} + \mathrm{IQR}_f^{(1)}\right)}, \tag{1}$$

TABLE I
DEMOGRAPHIC AND SLEEP CHARACTERISTICS OF THE STUDY POPULATION. VALUES ARE REPORTED AS MEAN ± STANDARD DEVIATION

Group	Sex (M/F)	Age (years)	N1+N2 (%)	N3 (%)	REM (%)	PSG duration (hh:mm)
HC (33)	22/11	51.6 ± 10.3	53.2 ± 13.9	19.0 ± 8.8	15.0 ± 6.0	$07:37 \pm 01:16$
iRBD (34)	18/16	56.6 ± 13.1	53.3 ± 10.7	16.6 ± 7.4	16.5 ± 6.4	$07:23 \pm 01:17$

TABLE II
HRV FEATURES BY DOMAIN AND AGGREGATION METRICS

Domain	Features
Time-domain	SDNN, RMSSD
Frequency-domain	VLF, HF, TP
Non-linear	SD2, SampEn, LZC, DFA α_1
Aggregation	Median, SD, Skewness, P_{90}

where $\tilde{x}_f^{(c)}$ is the median of feature f in class c and $\mathrm{IQR}_f^{(c)}$ the corresponding interquartile range. Features are sorted in descending order of SI_f.

2) Redundancy Pruning: The ranked list is walked greedily: each candidate feature is discarded if its absolute Spearman correlation with any already-accepted feature exceeds $|\rho_s| > 0.80$; otherwise it is accepted. The walk terminates after M features have been accepted or the list is exhausted.

E. Training and Evaluation Pipeline

Classification follows a nested Leave-One-Subject-Out (LOSO) scheme.

The additive structure of the EBM is central to the interpretability of the pipeline: the model expresses the log-odds of the positive class as

$$g(x) = \beta_0 + \sum_{j=1}^{p} f_j(x_j), \tag{2}$$

where each f_j is a univariate shape function learned through an ensemble of shallow decision trees. Because every prediction decomposes into a sum of individual feature contributions, no post-hoc approximation is required. This decomposability motivates the feature selection and hyperparameter choices described below.

1) Outer-Fold Feature Selection: Within each outer fold the two-phase filter is not applied to the full training set directly. Instead, the $N-1$ training subjects are split into $K_{\mathrm{fs}} = 3$ stratified folds and the ranking-pruning procedure is executed on the inner-training portion of each split only. Features are ranked first by selection frequency across splits, then by mean rank position to break ties. The top M features form the outer fold's consensus ranking.

2) Hyperparameter Optimisation: Optuna [18] with a Tree-structured Parzen Estimator (TPE, 10 random start-up trials) searches over 50 trials within each outer fold. Each trial proposes four EBM hyperparameters, listed in Table III, together with the number of features $n_{\mathrm{feat}} \in [1,\ M]$. The nested LOSO pipeline was repeated with ten random seeds to assess the robustness of the selected hyperparameters, feature subset, and final performance.

TABLE III
EBM HYPERPARAMETER CONFIGURATION

Group	Parameter	Value
Fixed	`early_stopping_rounds`	200
	`min_samples_leaf`	2
	`validation_size`	0.10
Tuned	`learning_rate`	{0.005, 0.01, 0.02}
	`smoothing_rounds`	{50, 75, 100}
	`max_leaves`	{2, 3}
	`interactions`	{0, 2}
	`n_features`	[2, 10]

Trials are evaluated via 10-fold stratified cross-validation. For each inner fold the ranking-pruning procedure is recomputed independently on its own inner-training partition, so that feature rankings are never influenced by inner-validation subjects. The first n_{feat} features of each inner fold's list are used to train and evaluate the EBM.

The optimisation objective is a Lower Confidence Bound (LCB) score:

$$\mathrm{LCB} = \bar{s} - 0.5\,\sigma_s, \tag{3}$$

where $\bar{s}$ is the mean accuracy across inner folds, σ_s its standard deviation, and $\kappa = 0.5$ is a penalisation coefficient set empirically to balance exploitation and stability across folds.

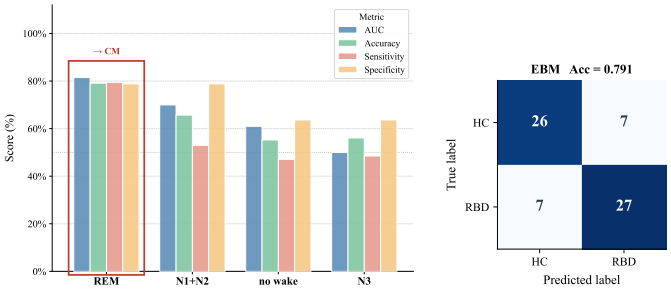
3) Final Model: The best Optuna trial determines the optimal hyperparameters and the final feature count. This feature count selects the top features from the consensus ranking computed on the $N-1$ training subjects. A single EBM is trained on all $N-1$ subjects using these specific features and hyperparameters. This final model predicts the label and class probability for the held-out test subject. To contextualise the EBM performance, Logistic Regression (LR) and Decision Tree (DT) baselines were evaluated across all four sleep stages under the same nested LOSO protocol, feature-selection pipeline, and Optuna-based hyperparameter optimisation.

III. RESULTS

A. Classification Performance

Figure 2(a) shows the EBM classification performance across the four sleep stages. Across ten random seeds, the

REM stage yielded the best performance, with mean AUC = 0.82 ± 0.02 and mean accuracy = 0.75 ± 0.04. LR and DT baselines, evaluated across all stages, also peaked at REM (seed 42: LR: AUC = 0.61, accuracy = 0.63; DT: AUC = 0.60, accuracy = 0.63), confirming REM as the most informative stage across all classifiers. Consequently, the feature evaluation and shape function analysis concentrated exclusively on the REM stage. Figure 2(b) provides the confusion matrix for the REM stage for the same representative run. For brevity, all subsequent detailed analyses refer to the representative run obtained with seed 42.



(a) (b)

Fig. 2. (a) Classification performance (AUC, accuracy, sensitivity, specificity) across sleep stages. (b) Confusion matrix for the REM stage.

B. Features and Hyperparameters Stability

Because each LOSO fold operates on a slightly different training set ($N-1$ subjects), both the consensus feature ranking and the Optuna-selected hyperparameters may vary across folds. Monitoring this variability is essential: a model whose feature set changes drastically from fold to fold offers weaker physiological interpretability than one that converges on a stable subset. Figure 3 shows the percentage of folds that include each feature within the final n subset, restricted to a minimum selection frequency of 10%. Four features achieved a selection frequency in over 85% of cases: `RMSSD_skew`, `SDNN_skew`, `SampEn_std`, and `DFA_alpha1_median`. This high frequency confirms the consensus method selects stable features across all leave-one-out folds.

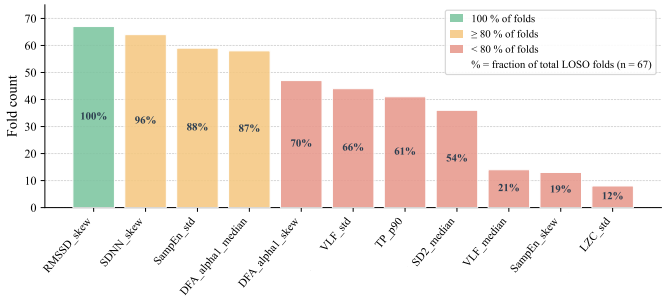


Fig. 3. Feature selection stability across LOSO folds. Each bar indicates in how many of the N folds a given feature was included in the final model.

Table IV reports the distribution of optimal hyperparameters selected by Optuna across the 67 outer folds.

TABLE IV
EBM HYPERPARAMETER DISTRIBUTION ACROSS 67 LOSO FOLDS (REM STAGE)

Parameter	Modal value	Frequency
`learning_rate`	0.005	49%
`smoothing_rounds`	75	43%
`max_leaves`	3	88%
`interactions`	0	88%
`n_features`	8	median = 8, range [3, 10]

To evaluate the statistical differences between HC and iRBD, statistical tests were performed on the four most frequently selected features. The normality of the feature distributions was first assessed using the Shapiro–Wilk test. As the distributions were found to be non-normal, the Mann–Whitney U test was applied to compare the two groups. The effect size was calculated using the rank-biserial correlation. The significance level was set at $\alpha = 0.05$; all four features were statistically significant in differentiating HC and iRBD. Figure 4 visualises the class-wise distributions of these features, together with the corresponding significance levels and effect size values.

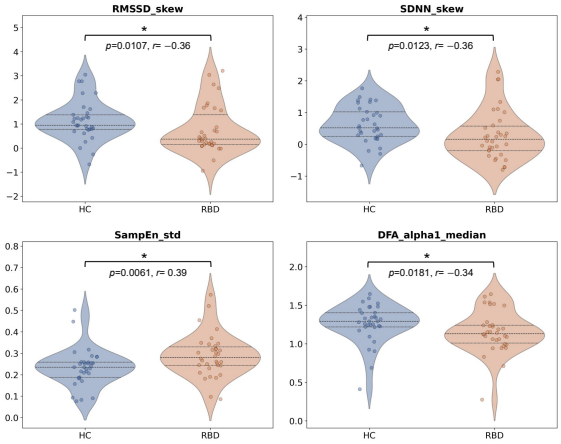


Fig. 4. Violin plots of HC and iRBD for the four most stable features. The p-value (p) and effect size (r) are reported; * denotes a statistically significant difference between groups.

1) Mean Additive Scores: For each subject the EBM assigns a log-odds contribution per feature; averaging these scores by true class reveals how each feature separates HC from iRBD at the population level. Figure 5 reports this analysis for the four core features.

`RMSSD_skew` exhibits the strongest bipolar separation: it contributes on average -0.8 log-odds for HC subjects and $+0.5$ log-odds for iRBD subjects, actively pushing the model toward the correct class in both groups. `DFA_alpha1_median` shows a similar symmetric pattern with smaller magnitude. `SampEn_std`, by contrast, displays an asymmetric profile: a strong negative contribution for HC but a near-zero contribution for iRBD, acting primarily as

a marker of healthy autonomic regularity rather than as an indicator of disease. SDNN_skew contributes weakly in both classes, yet is selected in 96% of the folds, indicating a small but generalisable correction.

It is worth noting that TP_p90 reaches the highest bipolar magnitude among the auxiliary features (± 0.2 log-odds), yet it is selected in only 61% of the folds. This combination of strong discriminative power and low selection stability suggests specialisation on a patient subset whose composition varies across folds, in contrast with the weak but almost global contribution of SDNN_skew.

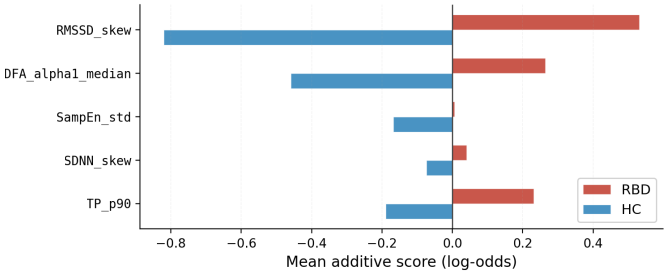


Fig. 5. Mean additive EBM score as log-odds per class in REM stage.

2) Shape Functions: While mean scores summarise each feature's average contribution, the shape functions $f_j(x_j)$ reveal how it varies with the feature value. Figure 6 plots, for each core feature, the median curve across LOSO folds (solid line) and the interquartile range (shaded band). Reading both representations jointly is informative: the shape function encodes the mapping rule learned by the model, while the mean score reflects its actual impact given the distribution of patients in the cohort.

a) RMSSD_skew: Values between approximately 0.75 and 1.6 produce strong negative scores ($f_j \approx -2.0$), assigning the subject to HC. Outside this range the curve shifts toward iRBD ($f_j \approx +0.5$). The IQR band widens at the extremes but remains on one side of zero, preserving a consistent direction across folds. The mean scores (-0.8 for HC, $+0.5$ for iRBD) confirm dense patient representation in the respective discriminative segments.

b) SDNN_skew: The curve decreases monotonically with a zero-crossing near 0.45: below this threshold it favours iRBD ($f_j \approx +1.0$), above it HC ($f_j \approx -0.75$). Mean scores are weak in both classes, indicating that most subjects cluster near the zero-crossing; nevertheless, the feature is selected in 96% of folds, confirming a small but consistently useful correction.

c) SampEn_std: A zero-crossing near 0.27 separates two regimes: below it the curve favours HC ($f_j \approx -0.6$); above it the score rises steeply toward iRBD ($f_j \approx +1.5$), the highest single-feature magnitude observed. Mean scores show that this feature acts primarily as an HC marker, as only a few subjects activate the steep positive segment, consistent with the wide IQR band in that region.

d) DFA_alpha1_median: Values between approximately 1.23 and 1.50 produce a strong negative score ($f_j \approx -2.5$ at the trough), clearly identifying HC. Outside this interval the curve rises toward iRBD, stabilising near $f_j \approx +1.0$ for values above 1.50 and below 0.90. Below 0.90 the IQR band spans both positive and negative predictions, reflecting complete model uncertainty in that region.

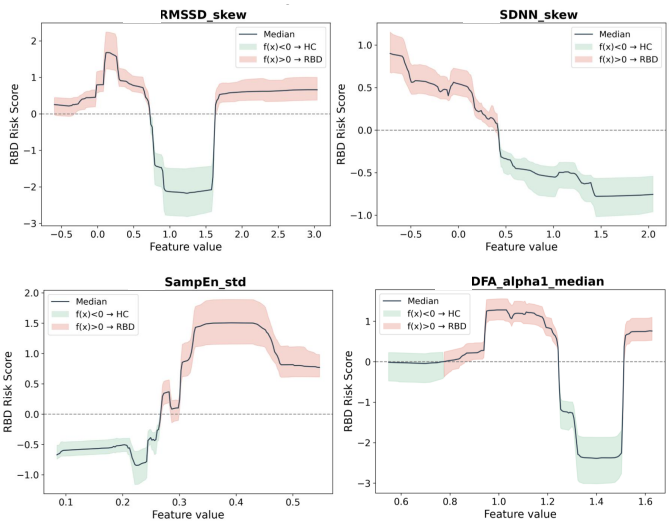


Fig. 6. EBM shape functions for the four most stable features (REM stage). Solid line: median $f_j(x_j)$ across LOSO folds; shaded band: interquartile range. Green: $f_j < 0$ (toward HC); red: $f_j > 0$ (toward iRBD).

IV. Discussion

A. Interpretable Autonomic Signatures of iRBD During REM Sleep

The central finding is that a stable set of HRV features extracted from the REM stage and classified by an EBM achieved a mean AUC of 0.82 ± 0.02 and a mean accuracy of 0.75 ± 0.04 in distinguishing iRBD from age-matched HC, with every prediction fully decomposed into additive shape functions. The EBM outperformed both LR and DT baselines, indicating that its choice is motivated not only by glass-box interpretability but also by stronger predictive performance. The superiority of REM over N1+N2, N3, and no-wake reflects the known pathophysiology: in healthy subjects REM is characterised by pronounced sympathetic surges and strong short-term fractal correlations, whereas early Lewy body infiltration of brainstem nuclei and sympathetic ganglia attenuates this dynamic in iRBD [5], [6].

Four features were selected in more than 85% of the 67 LOSO folds. All four were also statistically significant between groups with effect sizes above 0.3, and this agreement between cross-validation stability and statistical significance supports their relevance as autonomic markers of iRBD. The RMSSD_skew shape function assigns strong negative log-odds ($f_j \approx -2.0$) to the healthy range (0.75-1.6), flagging flattened or erratic distributions outside it. The DFA_alpha1_median identifies a healthy window (1.2-1.5) with a deep trough ($f_j \approx -2.5$), penalising values outside it.

B. Comparison with the Literature

Direct comparison with prior work is difficult. A recent systematic review [19] confirms severe methodological inhomogeneity across HRV-based RBD studies: the majority evaluate cardiac autonomic control using daytime tests rather than nocturnal PSG, and those that do analyse sleep data rarely stratify by stage. Machine learning approaches to REM-related parasomnias [10], [11] achieve competitive classification rates but rely on multi-channel setups and opaque algorithms (e.g., SVM), precluding direct inspection of learned decision rules. The present work occupies a distinct position by restricting inputs to single-lead ECG during individual sleep stages and employing an EBM whose additive structure yields exact physiological thresholds, deliberately prioritising transparency over maximal predictive performance.

C. Clinical and Translational Implications

Regulatory frameworks for medical AI require auditable models that support human oversight [13]. The EBM meets this requirement by providing exact, globally consistent feature contributions inspectable for any individual prediction. Beyond the clinic, single-lead ECG is recordable by consumer wearables, opening a pathway toward home-based screening for prodromal synucleinopathy and longitudinal monitoring of autonomic progression in neuroprotective trials.

D. Limitations and Future Directions

The dataset comprises 67 subjects from a single centre; validation on multi-site cohorts is necessary to establish generalisability. The sample size also limits subgroup and fairness analyses. Future work should also assess whether window-level approaches, such as multiple-instance learning or majority voting, better capture the paroxysmal nature of iRBD than fixed-length subject-level aggregation. Translation to wearable ECG introduces noisier RR estimates, absence of manual hypnograms, and potential domain shift. Future work should assess robustness to sensor noise and automated sleep staging, integrate additional PSG signals (EEG, EMG), extend the classification to calibrated conversion-risk scores, and investigate night-to-night variability of the identified HRV biomarkers.

V. Conclusion

This study demonstrated that HRV features extracted from single-lead ECG during sleep, computed separately for each sleep stage and aggregated via summary statistics, can reliably distinguish iRBD from HC when classified by an EBM. The REM stage proved most informative, and four core features were consistently selected across LOSO folds, confirming their potential as reproducible digital biomarkers. The additive structure of the EBM enables fully transparent predictions, offering an interpretable and physiologically grounded framework for automated iRBD screening.

References

[1] GBD 2016 Neurology Collaborators, "Global, regional, and national burden of neurological disorders, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016," *Lancet Neurology*, vol. 18, no. 5, pp. 459–480, 2019.

[2] A. Iranzo, E. Tolosa, and G. al., "Neurodegenerative disease status and post-mortem pathology in idiopathic rapid-eye-movement sleep behaviour disorder: an observational cohort study," *Lancet Neurology*, vol. 12, no. 5, pp. 443–453, 2013.

[3] R. B. Postuma and G. al., "Quantifying the risk of neurodegenerative disease in idiopathic REM sleep behavior disorder," *Neurology*, vol. 72, no. 15, pp. 1296–1300, 2009.

[4] American Academy of Sleep Medicine, *International Classification of Sleep Disorders*, 3rd ed. Darien, IL: American Academy of Sleep Medicine, 2014.

[5] E. E. Benarroch, "Brainstem in multiple system atrophy: clinicopathological correlations," *Cell and Molecular Neurobiology*, vol. 23, no. 4–5, pp. 519–526, 2003.

[6] R. B. Postuma and L. al., "Potential early markers of parkinson disease in idiopathic rem sleep behavior disorder," *Neurology*, vol. 66, no. 6, pp. 845–851, 2006.

[7] 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*, vol. 93, no. 5, pp. 1043–1065, 1996.

[8] V. K. Somers and D. al., "Sympathetic-nerve activity during sleep in normal subjects," *New England Journal of Medicine*, vol. 328, no. 5, pp. 303–307, 1993.

[9] K. Umetani and S. al., "Twenty-four hour time domain heart rate variability and heart rate: relations to age and gender over nine decades," *Journal of the American College of Cardiology*, vol. 31, no. 3, pp. 593–601, 1998.

[10] M. Cesari and I. Rechichi, "Automatic and machine learning methods for detection and characterization of rem sleep behavior disorder," in *Handbook of AI and Data Sciences for Sleep Disorders*. Springer, 2024, pp. 197–217.

[11] 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," *International Journal of Environmental Research and Public Health*, vol. 19, no. 1, p. 248, Dec. 2021. [Online]. Available: https://www.mdpi.com/1660-4601/19/1/248

[12] M. Salsone and Q. et al., "A machine learning approach for detecting idiopathic rem sleep behavior disorder," *Diagnostics*, vol. 12, no. 11, p. 2689, 2022.

[13] U.S. Food and Drug Administration, "Artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD) action plan," https://www.fda.gov/media/145022/download, 2021.

[14] 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, 2019.

[15] Y. Lou and C. et al., "Intelligible models for classification and regression," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2012, pp. 150–158.

[16] H. Nori and J. et al., "InterpretML: A unified framework for machine learning interpretability," 2019.

[17] D. Makowski and T. P. et al., "NeuroKit2: A python toolbox for neurophysiological signal processing," *Behavior Research Methods*, vol. 53, no. 4, pp. 1689–1696, feb 2021. [Online]. Available: https://doi.org/10.3758/s13428-020-01516-y

[18] T. Akiba and e. a. Sano, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 2623–2631.

[19] C. Gino and D. V. et al., "Role of heart rate variability analysis in differentiating patients with idiopathic REM sleep behaviour disorders from healthy subjects. A systematic review with meta-analysis," *Sleep Medicine Reviews*, vol. 84, p. 102160, Dec. 2025. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1087079225001133