RCMOS – Revista Científica Multidisciplinar O Saber. ISSN: 2675-9128. São Paulo - SP.
Ano VII, v.1 2026 | submissão: 23/07/2026 | aceito: 26/07/2026 | publicação: 29/07/2026

Automatic Sleep Stage Classification with Subject-Wise Validation: Comparison of Random Forest, XGBoost, and SVM on the Sleep-EDFx Dataset

Classificação Automática dos Estágios do Sono com Validação por Indivíduo: Comparação entre Random Forest, XGBoost e SVM no Conjunto de Dados Sleep-EDFx

Clasificación Automática de las Etapas del Sueño con Validación por Sujeto: Comparación de Random Forest, XGBoost y SVM en el Conjunto de Datos Sleep-EDFx

Francisco Hugo César Pereira Vale – Institute of Informatics (INF), Federal University of Goiás (UFG), Goiânia, Brazil, franciscovale@discente.ufg.br
Ronaldo Martins da Costa – Institute of Informatics (INF), Federal University of Goiás (UFG), Goiânia, Brazil, ronaldocosta@ufg.br

Abstract: This paper presents a reproducible pipeline for automatic classification of sleep stages (Wake W, N1, N2, N3, and REM) using single-channel EEG (Fpz-Cz) from the public Sleep-EDFx dataset (PhysioNet), comprising 100 subjects and 175 recordings from two distinct subsets: Sleep Cassette (SC, 78 healthy subjects, home recordings) and Sleep Telemetry (ST, 22 subjects with mild sleep onset difficulty, placebo nights only). A total of 17 features were extracted per 30-second epoch across two domains: time-domain statistics (mean, variance, standard deviation, skewness, kurtosis, and Hjorth parameters) and frequency-domain features via Welch spectral estimation (absolute and relative power in the Delta, Theta, Alpha, and Beta bands, plus the Delta/Alpha spectral ratio). Three classifiers of fundamentally different algorithmic natures were compared (Random Forest, XGBoost, and Support Vector Machine with RBF kernel) under Leave-One-Group-Out (LOGO) cross-validation, preventing any data leakage between individuals. Experiments were conducted in three scenarios: SC alone, ST alone, and the combined dataset. Random Forest achieved the best overall performance, with Cohen's Kappa of 0.587 on SC and 0.583 on the combined dataset. The convergence of all three classifiers to a similar Kappa range (0.53 to 0.59) indicates that the performance ceiling is primarily determined by the feature set and the inherent physiological overlap between stages, rather than by the choice of classifier.

Keywords: EEG signal processing. Machine learning. Polysomnography. Sleep staging.

Resumo: Este artigo apresenta um pipeline reprodutível para a classificação automática dos estágios do sono (Vigília W, N1, N2, N3 e REM) utilizando EEG de canal único (Fpz-Cz) do conjunto de dados público Sleep-EDFx (PhysioNet), composto por 100 indivíduos e 175 registros de dois subconjuntos distintos: Sleep Cassette (SC, 78 indivíduos saudáveis, registros domiciliares) e Sleep Telemetry (ST, 22 indivíduos com dificuldade leve para iniciar o sono, somente noites com placebo). Foram extraídas 17 características por época de 30 segundos em dois domínios: estatísticas no domínio do tempo (média, variância, desvio padrão, assimetria, curtose e parâmetros de Hjorth) e características no domínio da frequência por meio da estimativa espectral de Welch (potência absoluta e relativa nas bandas Delta, Teta, Alfa e Beta, além da razão espectral Delta/Alfa). Três classificadores de naturezas algorítmicas fundamentalmente diferentes foram comparados (Random Forest, XGBoost e Máquina de Vetores de Suporte com kernel RBF) sob validação cruzada Leave-One-Group-Out (LOGO), impedindo qualquer vazamento de dados entre indivíduos. Os experimentos foram conduzidos em três cenários: somente SC, somente ST e o conjunto combinado. O Random Forest obteve o melhor desempenho geral, com Kappa de Cohen de 0,587 no SC e 0,583 no conjunto combinado. A convergência dos três classificadores para uma faixa de Kappa semelhante (0,53 a 0,59) indica que o teto de desempenho é determinado principalmente pelo

conjunto de características e pela sobreposição fisiológica inerente entre os estágios, e não pela escolha do classificador.

Palavras-chave: Processamento de sinais de EEG. Aprendizado de máquina. Polissonografia. Estadiamento do sono.

Resumen: Este artículo presenta un pipeline reproducible para la clasificación automática de las etapas del sueño (Vigilia W, N1, N2, N3 y REM) utilizando EEG de un solo canal (Fpz-Cz) del conjunto de datos público Sleep-EDFx (PhysioNet), compuesto por 100 sujetos y 175 registros de dos subconjuntos distintos: Sleep Cassette (SC, 78 sujetos sanos, registros domiciliarios) y Sleep Telemetry (ST, 22 sujetos con dificultad leve para conciliar el sueño, únicamente noches con placebo). Se extrajeron 17 características por época de 30 segundos en dos dominios: estadísticas en el dominio temporal (media, varianza, desviación estándar, asimetría, curtosis y parámetros de Hjorth) y características en el dominio de la frecuencia mediante estimación espectral de Welch (potencia absoluta y relativa en las bandas Delta, Theta, Alfa y Beta, además de la razón espectral Delta/Alfa). Se compararon tres clasificadores de naturalezas algorítmicas fundamentalmente diferentes (Random Forest, XGBoost y Máquina de Vectores de Soporte con kernel RBF) mediante validación cruzada Leave-One-Group-Out (LOGO), evitando cualquier fuga de datos entre sujetos. Los experimentos se realizaron en tres escenarios: solo SC, solo ST y el conjunto combinado. Random Forest obtuvo el mejor desempeño general, con Kappa de Cohen de 0,587 en SC y 0,583 en el conjunto combinado. La convergencia de los tres clasificadores hacia un intervalo de Kappa similar (0,53 a 0,59) indica que el techo de desempeño está determinado principalmente por el conjunto de características y la superposición fisiológica inherente entre las etapas, más que por la elección del clasificador.

Palabras clave: Procesamiento de señales EEG. Aprendizaje automático. Polisomnografía. Estadificación del sueño.

1. Introduction

Sleep is a fundamental physiological process for organismal homeostasis, directly associated with memory consolidation [1], metabolic regulation, immune function, and cardiovascular health [2]. Sleep disorders affect a significant portion of the world population and constitute a risk factor for conditions such as hypertension, type 2 diabetes, depression, and neurodegenerative diseases [2]. Clinical diagnosis of these conditions depends on the precise identification of sleep stages (Wake W, N1, N2, N3, and REM) from polysomnographic (PSG) recordings.

The current gold standard for this diagnosis is polysomnography (PSG) with manual staging performed by trained specialists, who visually analyze the electroencephalogram (EEG), electrooculogram (EOG), and electromyogram (EMG), segmented into 30-second scoring epochs (in the clinical PSG sense [3], not to be confused with training epochs in neural network optimization) and classified according to the criteria of the American Academy of Sleep Medicine (AASM) [3]. Although clinically established, this process presents significant

limitations: it is time-consuming, costly, and subject to inter-rater variability. Meta-analysis studies indicate that Cohen's Kappa between human specialists for global staging is approximately 0.76, but drops to only 0.24 for the N1 stage, showing that even experienced annotators substantially disagree on transitional stages [4]. This intrinsic variability represents a natural ceiling for any supervised automatic system, since models are trained on annotations that already contain this agreement noise [5].

Given this scenario, automatic sleep stage classification through machine learning has emerged as a clinically and technologically relevant research area. Approaches based on handcrafted features (manual extraction of spectral and statistical EEG characteristics) followed by classical classifiers such as Random Forest, SVM, and XGBoost demonstrate competitive performance and interpretability advantages over deep learning architectures, especially when the number of available subjects is limited [6], [7]. At the same time, deep neural network approaches such as 1D CNN, LSTM, and Transformers have reported Kappa values above 0.80 on large-scale datasets with multiple PSG channels [8], [9].

This work evaluates automatic sleep stage classification using the public Sleep-EDFx dataset from PhysioNet [10], comprising the Sleep Cassette (SC) and Sleep Telemetry (ST) subsets. Three classifiers of different algorithmic natures are compared (Random Forest, XGBoost, and SVM) using 17 features extracted from single-channel EEG (Fpz-Cz), with rigorous Leave-One-Group-Out (LOGO) validation.

2. Methodology

2.1 Dataset

The Sleep-EDFx is a public dataset available on PhysioNet [10], widely used as a benchmark in the automatic sleep staging literature [6], [9]. Two subsets were used: Sleep Cassette (SC), with 78 healthy subjects aged 25 to 101 years, without sleep-related medication, totaling 153 recordings of approximately 20 hours performed in a home environment; and Sleep Telemetry (ST), with 22 subjects with mild sleep onset difficulty, recorded in a hospital setting, using exclusively the Placebo condition nights (22 recordings of approximately 9 hours), identified from the official PhysioNet metadata spreadsheet ST-subjects.xls. Annotations follow the Rechtschaffen & Kales (R&K) convention, with stages N3 and N4 merged into a single N3 class according to the current AASM convention [3], resulting in five classes: W, N1, N2, N3, and REM.

RCMOS – Revista Científica Multidisciplinar O
Saber. ISSN: 2675-9128. São Paulo - SP.
Ano VII, v.1 2026 | submissão: 23/07/2026 | aceito: 26/07/2026 | publicação: 29/07/2026

Ethics statement: This study used exclusively publicly available secondary data from the Sleep-EDFx dataset (PhysioNet). No direct interaction with human subjects was conducted. The original data collection was performed in accordance with the Declaration of Helsinki and approved by the local ethics committees of the responsible institutions. Data are fully anonymized, with subjects identified only by numeric codes. Use for research purposes is permitted under the PhysioNet Credentialed Health Data License. No additional ethics committee approval was required for this work.

Figure 1 illustrates the steps of the proposed pipeline, from reading the EDF files to the final evaluation of the classifiers.

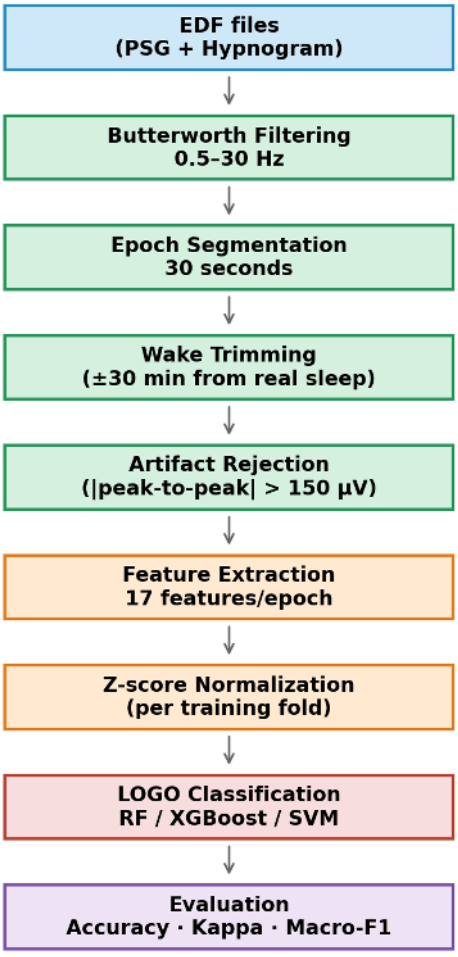

Figure 1. Automatic sleep stage classification pipeline. Preprocessing steps are applied individually to each recording before feature extraction.

2.2 Preprocessing

EEG signals from the Fpz-Cz channel were filtered with a 4th-order Butterworth bandpass filter (0.5 to 30 Hz), preserving the Delta, Theta, Alpha, and Beta bands relevant for sleep staging. Epochs with peak-to-peak amplitude exceeding 150 µV were discarded through artifact rejection. To reduce the artificial imbalance of the Wake class, a wake trimming procedure was applied: only epochs within 30 minutes before the first real sleep epoch (N1, N2, N3, or REM) and 30 minutes after the last were retained, discarding the long wake periods at the edges of the recording.

Figure 2 illustrates the hypnogram of a representative subject after wake trimming, showing the typical cyclic structure of human sleep, with alternation between N1, N2, N3, and REM throughout the night, and predominance of the N2 stage [3].

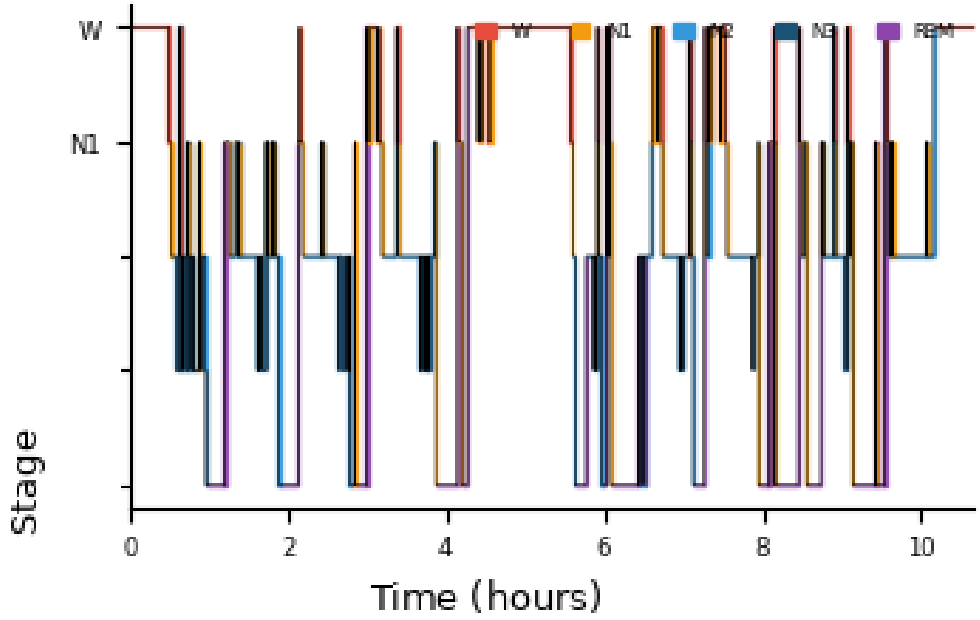

Figure 2. Example hypnogram after wake trimming. Each color represents a sleep stage: Wake (W), N1, N2, N3, and REM.

2.3 Feature Extraction

A total of 17 features were extracted per 30-second epoch from the Fpz-Cz EEG channel, organized into two domains. In the time domain: mean, variance, standard deviation, skewness, kurtosis, and the three Hjorth parameters (activity, mobility, and complexity). In the frequency domain, via Welch spectral estimation: absolute and relative power in the Delta (0.5 to 4 Hz), Theta (4 to 8 Hz), Alpha (8 to 12 Hz), and Beta (12 to 30 Hz) bands, plus the Delta/Alpha spectral ratio, a relevant indicator for distinguishing deep sleep from wakefulness.

2.4 Classifiers and Validation

Three classifiers of fundamentally different algorithmic natures were evaluated: Random Forest (bagging ensemble, 300 estimators, balanced class weights), XGBoost (boosting ensemble, 300 estimators, maximum depth 6), and Support Vector Machine with RBF kernel, with class balancing applied in all models. For the SVM, due to the quadratic computational complexity of the RBF kernel on large datasets, each fold was trained on a stratified subsample of up to 8,000 epochs, preserving the original class proportions, a standard practice in the literature for kernel-based classifiers at scale [7]. Features were normalized using z-score standardization fitted exclusively on the training fold, preventing statistical information leakage to the test set. Validation followed the Leave-One-Group-Out (LOGO) strategy, in which each fold excludes one subject entirely for testing and trains on all others, ensuring that no epoch from the same subject appears simultaneously in training and test sets. The three classifiers were evaluated in three independent scenarios: Sleep Cassette alone (78 folds), Sleep Telemetry Placebo alone (22 folds), and combined dataset (100 folds).

3. Results

Table 1 presents the consolidated results of the three classifiers across the three evaluation scenarios, expressed in Accuracy, Macro-F1, and Cohen's Kappa.

Table 1. Classifier Performance by Scenario (LOGO)

Scenario	Classifier	Acc.	Macro-F1	Kappa
Cassette (n=78)	Random Forest	0.705	0.625	0.587
Cassette (n=78)	XGBoost	0.675	0.630	0.568
Cassette (n=78)	SVM (RBF)*	0.642	0.602	0.531
Telemetry (n=22)	Random Forest	0.687	0.596	0.548
Telemetry (n=22)	XGBoost	0.704	0.629	0.582
Telemetry (n=22)	SVM (RBF)*	0.673	0.605	0.554
Combined (n=100)	Random Forest	0.700	0.625	0.583
Combined (n=100)	XGBoost	0.673	0.631	0.567
Combined (n=100)	SVM (RBF)*	0.646	0.606	0.536

**SVM: stratified subsample of 8,000 epochs/fold.*

Random Forest achieved the best overall performance, with Cohen's Kappa of 0.587 in the Cassette scenario and 0.583 in the combined dataset. XGBoost presented competitive results, outperforming Random Forest in the Telemetry scenario (Kappa 0.582 vs. 0.548). The SVM

RCMOS – Revista Científica Multidisciplinar O
Saber. ISSN: 2675-9128. São Paulo - SP.
Ano VII, v.1 2026 | submissão: 23/07/2026 | aceito: 26/07/2026 | publicação: 29/07/2026

with RBF kernel, evaluated with a training subsample, showed lower accuracy but comparable Macro-F1, highlighting its sensitivity to the volume of available training data.

It is observed that combining the two subsets (SC + ST, 100 subjects) did not result in performance degradation relative to the isolated scenarios, maintaining Kappa in the range of 0.53 to 0.58 for all classifiers. This suggests that EEG sleep patterns between healthy subjects (Cassette) and subjects with mild sleep onset difficulty under placebo (Telemetry) are sufficiently homogeneous for a jointly trained classifier to generalize across both populations.

Among the 17 extracted features, spectral-domain features tend to be physiologically more informative for sleep staging, since each stage presents a characteristic band signature: delta waves predominate in deep sleep (N3), sleep spindles and K-complexes are markers of N2, and theta and alpha activity characterize N1 and wakefulness, respectively [3]. A quantitative analysis of the relative importance of features is planned as a subsequent step of this work.

Figure 3 presents the normalized confusion matrix of the Random Forest classifier in the combined scenario, aggregating all out-of-sample predictions from the 100 LOGO folds.

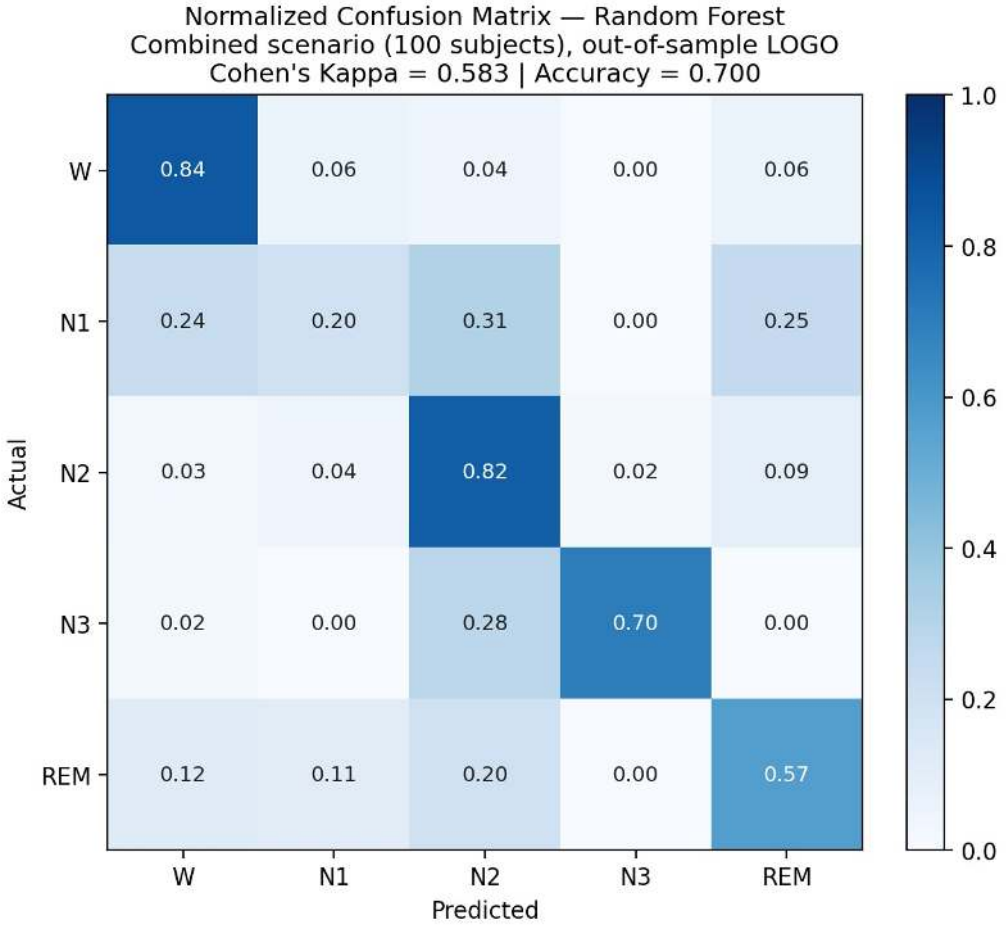

Figure 3. Normalized confusion matrix of Random Forest, combined scenario (100 subjects), aggregating all out-of-sample LOGO predictions. Values represent the proportion of real epochs classified into each predicted stage.

RCMOS – Revista Científica Multidisciplinar O
Saber. ISSN: 2675-9128. São Paulo - SP.
Ano VII, v.1 2026 | submissão: 23/07/2026 | aceito: 26/07/2026 | publicação: 29/07/2026

Figures 4, 5, and 6 present, respectively, the Accuracy, Cohen's Kappa, and Macro-F1 of the three classifiers across the three evaluated scenarios, enabling a direct visual comparison between algorithms.

Regarding Accuracy (Figure 4), Random Forest achieved the best results in the Cassette (0.705) and Combined (0.700) scenarios, while XGBoost outperformed the others in the Telemetry scenario (0.704). The SVM with RBF kernel showed consistently lower accuracy, reflecting the training subsampling adopted due to computational constraints.

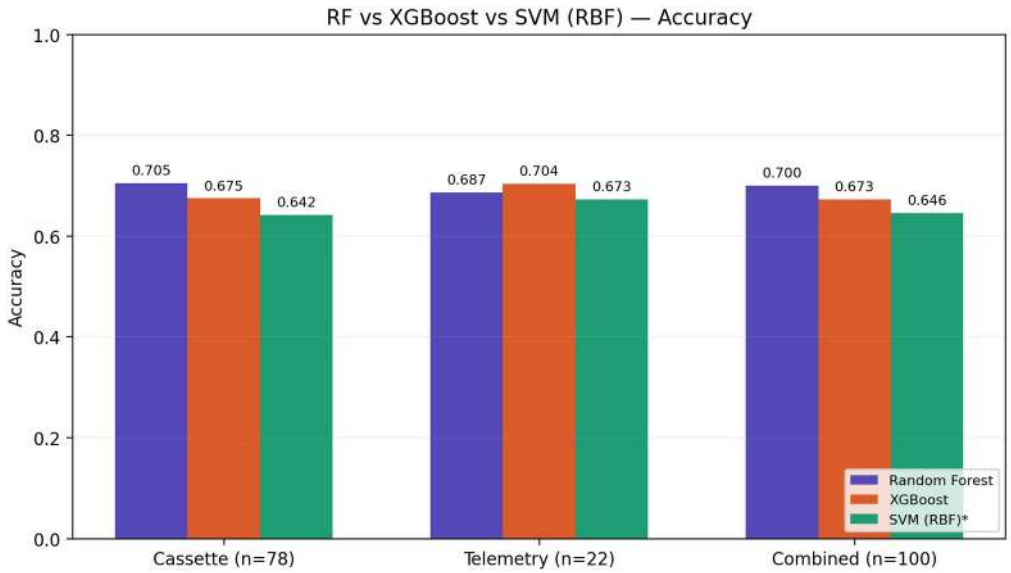

Figure 4. Classifier accuracy by scenario (LOGO). *SVM trained on a subsample of 8,000 epochs/fold.

Cohen's Kappa (Figure 5), the standard metric in sleep staging literature as it discounts the chance agreement effect, confirms the same pattern: Random Forest leads in Cassette (0.587) and Combined (0.583), XGBoost in Telemetry (0.582), and SVM maintains competitive performance given the training restriction (0.531 to 0.554). All values fall in the moderate to substantial range by the Landis & Koch scale, comparable to the inter-rater human agreement floor for individual stages reported in the literature [4].

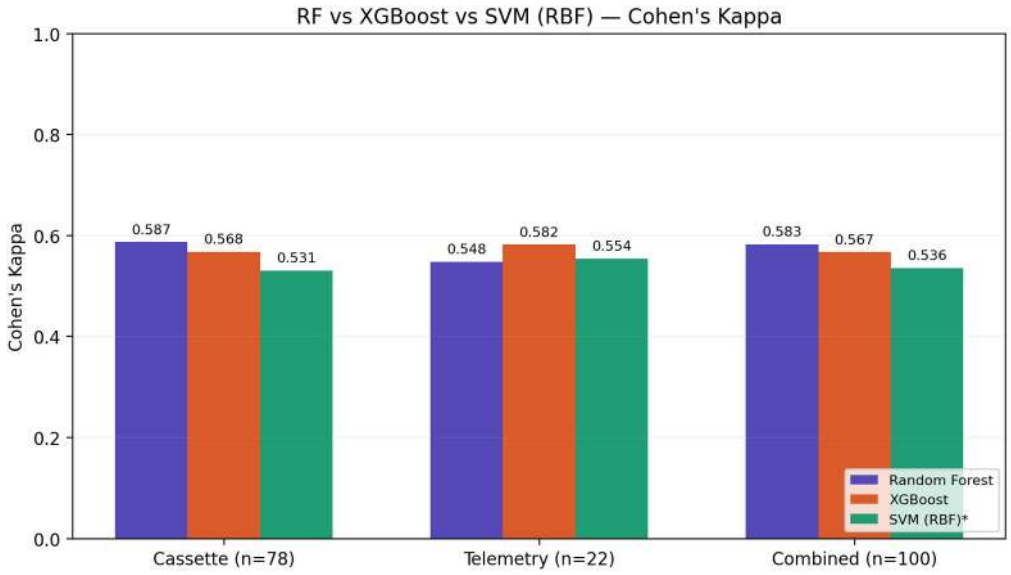

Figure 5. Cohen's Kappa of classifiers by scenario (LOGO).

Macro-F1 (Figure 6) reveals a distinct pattern: XGBoost presents the highest Macro-F1 across all three scenarios (0.630, 0.629, and 0.631), outperforming Random Forest, which leads in Accuracy and Kappa. This inversion indicates that XGBoost better classifies minority classes (especially N1 and N3), while Random Forest tends to concentrate correct predictions on majority classes (N2 and Wake). For clinical applications where the correct identification of all stages is equally relevant, Macro-F1 constitutes an indispensable complementary metric to Accuracy and Cohen's Kappa.

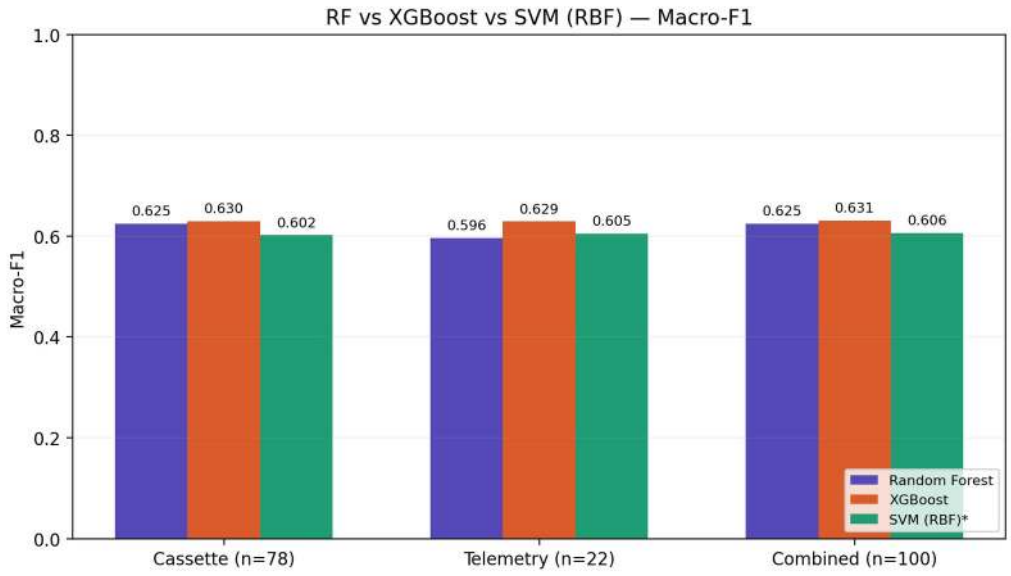

Figure 6. Macro-F1 of classifiers by scenario (LOGO).

Ano VII, v.1 2026 | submissão: 23/07/2026 | aceito: 26/07/2026 | publicação: 29/07/2026

4. Discussion

Cohen's Kappa values obtained (0.53 to 0.59) fall in the moderate to substantial range by the Landis & Koch scale, comparable to the floor of inter-rater human agreement reported for individual stages in the literature. In particular, the human Kappa for N1 is only 0.24 [4], the stage that also constitutes the main challenge for the automatic classifiers in this work. The fact that three algorithms of fundamentally different natures (bagging, boosting, and geometric boundary) converge to the same performance range suggests that the observed ceiling is a property of the feature set and the inherent physiological overlap between stages, rather than an artifact of the classifier choice.

The LOGO validation strategy adopted is more rigorous than the random k-fold frequently employed in the literature, as it eliminates any possibility of information leakage between epochs of the same subject, a common source of artificial optimism in works that partition by epoch rather than by subject [7].

The normalized confusion matrix of Random Forest in the combined scenario (Figure 3) highlights error patterns consistent with the automatic sleep staging literature. Stage N2 achieves the highest recognition rate (0.83), reflecting its high prevalence in the dataset and its characteristic electrophysiological markers. Stage REM shows good discrimination (0.70), with confusion mainly with N1 (0.14) and N2 (0.13). Stage N3 is frequently confused with N2 (0.57), an expected difficulty for aggregated band features in 30-second windows. Stage N1 is the most challenging (0.31), with errors distributed among W, N2, and REM [4].

5. Conclusion

This work presented a reproducible pipeline for automatic sleep stage classification using the Sleep-EDFx dataset from PhysioNet, with rigorous subject-wise validation (LOGO) and comparison of three classifiers of distinct natures. Results demonstrate that approaches based on handcrafted single-channel EEG features are viable and interpretable, achieving Cohen's Kappa of up to 0.587 with Random Forest, compatible with the literature for similar single-channel and LOGO validation configurations.

The choice of classical classifiers over deep learning architectures was deliberate: models such as Random Forest, XGBoost, and SVM present lower computational inference cost, a fundamental requirement for deployment on mobile devices and embedded systems with limited processing resources. This positions the evaluated classifiers as natural candidates for

integration in real-time sleep monitoring applications, where computational efficiency is as relevant as model accuracy.

As future work, it is proposed: (i) to develop a web application that makes the proposed pipeline available as a service, allowing users to upload EDF files and obtain automatic sleep stage classification with interactive hypnogram generation, democratizing access to automated staging without requiring local infrastructure; (ii) to expand features by incorporating EOG and EMG channels and temporal context between consecutive epochs; (iii) to conduct a systematic analysis of feature importance, identifying the most discriminative spectral bands and statistical parameters per stage, with the potential to reduce the feature set and optimize computational cost for mobile deployment; and (iv) to evaluate SVM with RBF kernel on the complete dataset per fold, using scalable implementations such as SGDClassifier with kernel approximation via the Nyström method, eliminating the need for subsampling.

References

DAULATABAD, V. et al. Systematic review: the impact of poor sleep quality on psychological and physical health. Journal of Heart Valve Disease, v. 30, p. 79–85, 2025.

FIORILLO, L.; MONACHINO, A.; VAN DER MEER, M.; WARNCKE, V.; SCHMIDT, M.; WENZ, F.; GRUN, R.; FARACI, F. Bridging AI and clinical practice: Integrating automated sleep scoring algorithm with uncertainty-guided physician review. arXiv preprint arXiv:2312.14996, 2023.

GOLDBERGER, A. L.; AMARAL, L. A. N.; GLASS, L.; HAUSDORFF, J. M.; IVANOV, P. Ch.; MARK, R. G.; MIETUS, J. E.; MOODY, G. B.; PENG, C.-K.; STANLEY, H. E. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation, v. 101, n. 23, p. e215–e220, 2000.

IBER, C.; ANCOLI-ISRAEL, S.; CHESSON, A.; QUAN, S. F. The AASM Manual for the Scoring of Sleep and Associated Events. Westchester, IL: American Academy of Sleep Medicine, 2007.

LEE, Y. J.; LEE, J. Y.; CHO, J. H.; CHOI, J. H. Interrater reliability of sleep stage scoring: a meta-analysis. Journal of Clinical Sleep Medicine, v. 18, n. 1, p. 193–202, 2022.

MASAD, I. S.; ALQUDAH, A.; QAZAN, S. Automatic classification of sleep stages using EEG signals and convolutional neural networks. PLOS ONE, v. 19, n. 1, p. e0297582, Jan. 2024.

SEKKAL, R. N.; BEREKSI-REGUIG, F.; RUIZ-FERNANDEZ, D.; DIB, N.; SEKKAL, S. Automatic sleep stage classification: From classical machine learning methods to deep learning. Biomedical Signal Processing and Control, v. 77, p. 103751, 2022.

VAN DER AAR, J. F.; VAN DEN ENDE, D. A.; VAN MEULEN, F. B.; VAN GILST, M. M. Deep transfer learning for automated single-lead EEG sleep staging with channel and population mismatches. Frontiers in Physiology, v. 14, Jan. 2024.

WEIGHALL, A.; KELLAR, I. Sleep and memory consolidation in healthy, neurotypical children and adults: a summary of systematic reviews and meta-analyses. Emerging Topics in Life Sciences, v. 7, n. 5, p. 513–524, Dec. 2023.

XU, X.; ZHANG, B.; XU, T.; TANG, J. An effective and interpretable sleep stage classification approach using multi-domain EEG and EOG features. Bioengineering, v. 12, n. 3, p. 286, 2025.