Feature Optimization for Automatic Sleep Classification with Random Forest

Francisco Hugo César Pereira Vale[1], Normando Perazzo Barbosa Souto[2], Ronaldo Martins da Costa[3]

[2]Federal University of Goiás (UFG), Goiânia, Brazil
(franciscovale@discente.ufg.br)
[2]University of the Federal District (UnDF), Brasília, Brazil,
[3]Federal University of Goiás (UFG), Goiânia, Brazil

Abstract: This work investigates feature selection based on Random Forest importance for automatic sleep stage classification (W, N1, N2, N3, and REM) using Sleep-EDFx (100 subjects, single-channel EEG Fpz-Cz). Subsets (top 5, 8, 10, 12, and 17 features) were evaluated under Leave-One-Subject-Out validation. The 8-feature model outperformed the full set: Cohen's Kappa 0.5925 and accuracy 70.68%, with 53% fewer features, indicating improved generalization and lower cost for mobile and PWA applications.

Keywords: sleep staging; Random Forest; feature selection; electroencephalogram; Sleep-EDFx.

INTRODUCTION

Sleep is an essential physiological process associated with memory consolidation, metabolic regulation, immune function, and cardiovascular health (WEIGHALL; KELLAR, 2023). Sleep disorders are a risk factor for hypertension, type 2 diabetes, depression, and neurodegenerative diseases (DAULATABAD et al., 2025). Clinical diagnosis depends on polysomnography (PSG) with manual staging according to AASM criteria (IBER et al., 2007), a time-consuming, costly process subject to inter-rater variability, Cohen's Kappa between specialists drops to only 0.24 for the N1 stage (LEE et al., 2022).

Approaches based on handcrafted EEG features combined with classical classifiers demonstrate competitive performance and superior interpretability over deep learning when the number of subjects is limited (XU et al., 2025; SEKKAL et al., 2022). However, most studies use fixed feature sets without systematically investigating the individual contribution of each feature. Feature selection is a fundamental step in machine learning pipelines, with direct impact on accuracy, generalization, and computational efficiency (GUYON; ELISSEEFF, 2003), an essential requirement for deployment on mobile devices, embedded systems, and Progressive Web Apps (PWA). This work systematically investigates the impact of feature selection by Random Forest importance on automatic sleep stage classification, evaluating growing subsets under LOSO validation on the Sleep-EDFx dataset.

MATERIAL AND METHODS

The Sleep-EDFx dataset (GOLDBERGER et al., 2000) was used in two subsets: Sleep Cassette (SC, 78 healthy subjects, 153 home recordings of ~20 hours) and Sleep Telemetry (ST, 22 subjects, Placebo nights only, 22 hospital recordings of ~9 hours), totaling 100 subjects, 175 recordings, and 178,344 epochs. Annotations follow the R&K convention with five classes: W, N1, N2, N3, and REM (IBER et al., 2007).

EEG signals from the Fpz-Cz channel were filtered with a 4th-order Butterworth bandpass filter (0.5–30 Hz), segmented into 30-second epochs, and subjected to artifact rejection (peak-to-peak amplitude > 150 μV) and wake trimming (±30 minutes from actual sleep onset). A total of 17 features were extracted per epoch: temporal statistical parameters (mean, variance, standard deviation, skewness, kurtosis, and Hjorth parameters) and spectral features via Welch (absolute and relative power in Delta, Theta, Alpha, and Beta bands, plus the Delta/Alpha ratio).

The importance of each feature was calculated using the Gini criterion of a Random Forest trained on a stratified subsample of 8,000 epochs. Growing subsets (top 5, 8, 10, 12, and 17) were evaluated with Random Forest (300 estimators, class_weight='balanced') under LOSO validation (100 folds), with z-score normalization per fold. The metrics, Accuracy, Macro-F1, and Cohen's Kappa , were calculated over all aggregated out-of-sample predictions from the 100 folds.

Figure 1 presents the complete importance ranking of the 17 features calculated by the Random Forest Gini criterion, showing the predominance of Hjorth Mobility and Skewness among the most discriminative features, followed by Delta and Beta spectral powers.

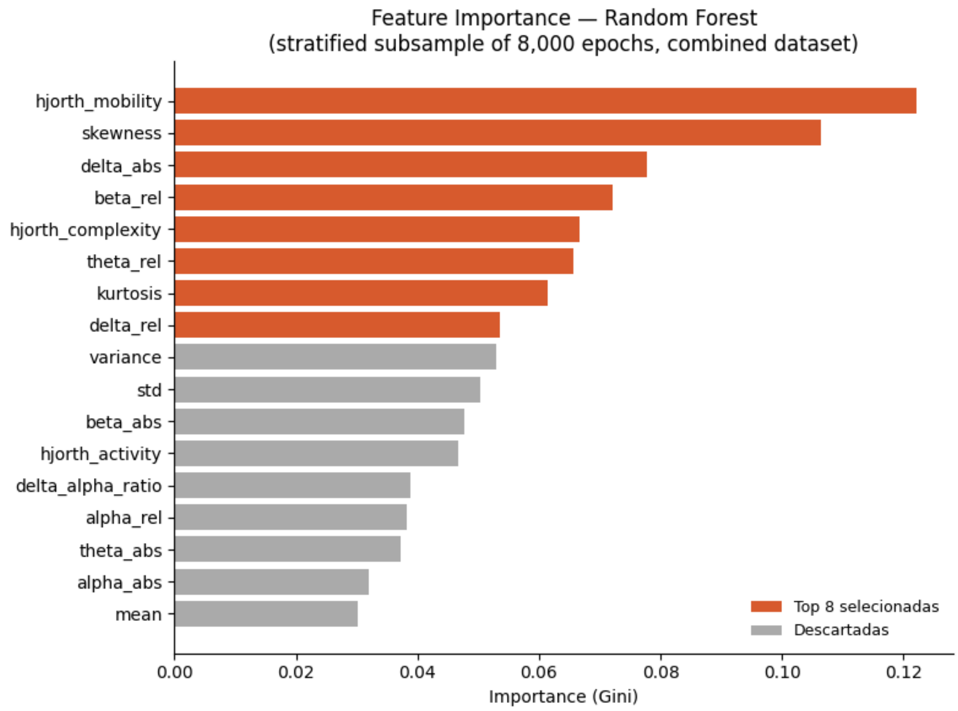

Figure 1: Relative importance of the 17 features calculated by the Random Forest Gini criterion (stratified subsample of 8,000 epochs, combined SC+ST dataset). The dashed line indicates the top 8 subset cutoff.

RESULTS AND DISCUSSION

Table 1 presents the Random Forest performance for each feature subset evaluated.

Table 1: Random Forest performance by feature subset (LOSO, 100 subjects)

Top N	Accuracy	Macro-F1	Cohen's Kappa
5	0,6782	0,5996	0,5543
8	0,7068	0,6298	**0,5925**
10	0,7037	0,6283	0,5879
12	0,7020	0,6269	0,5854
17 (full)	0,7008	0,6251	0,5834

The top 8 feature subset achieved the best performance across all metrics, with Cohen's Kappa of 0.5925, accuracy of 70.68%, and Macro-F1 of 0.6298, outperforming the 17-feature model (0.5834, 70.08%, and 0.6251). Beyond top 8, adding features progressively degrades all metrics, indicating that lower-importance features introduce noise rather than useful information.

The 8 optimal features are: Hjorth Mobility (0.122), Skewness (0.106), absolute Delta (0.078), relative Beta (0.072), Hjorth Complexity (0.067), relative Theta (0.066), Kurtosis (0.061), and relative Delta (0.054). The temporal domain dominates the most important positions, Hjorth Mobility captures differences in signal dynamics between stages (N3 presents slow and regular dynamics; REM and wakefulness, fast and irregular dynamics) and Skewness captures amplitude distribution asymmetries associated with the slow waves of N3. The spectral features Delta and Beta complement the set, capturing respectively the low-frequency energy of deep sleep and the high-frequency activity of wakefulness and REM.

Figure 2 illustrates the behavior of the three metrics as a function of the number of features used, highlighting the optimal point at top 8 and the progressive performance degradation with the addition of lower-importance features.

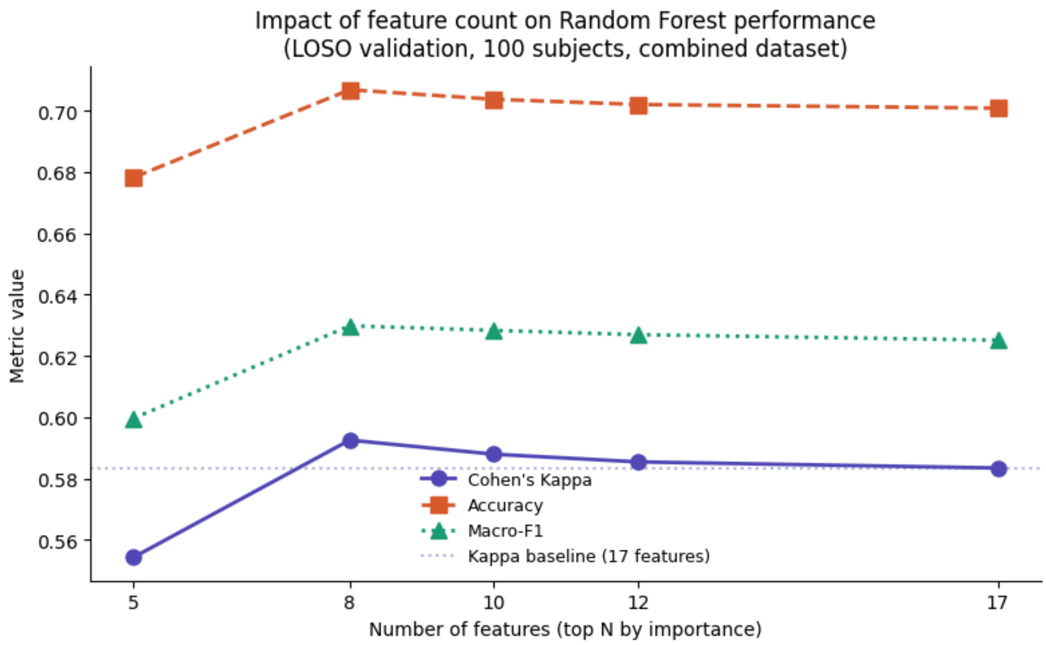

Figure 2: Impact of the number of features on Random Forest performance (Accuracy, Macro-F1, and Cohen's Kappa) under LOSO validation with 100 subjects. The horizontal dotted line indicates the Cohen's Kappa of the model with the full set of 17 features (baseline = 0.5834).

The 53% reduction in the number of features (from 17 to 8) represents an equivalent decrease in extraction cost and input vector size, without performance loss. For applications on mobile devices, embedded systems, and PWA, this result has a direct practical implication: the optimized model is lighter, faster, and more suitable for real-time deployment.

CONCLUSION

This work demonstrated that feature selection based on Random Forest importance reduces by 53% the feature set required for automatic sleep stage classification, with simultaneous gains across all metrics. The optimal 8-feature subset achieved Cohen's Kappa of 0.5925 and accuracy of 70.68%, outperforming the full 17-feature model. The most discriminative features combine temporal Hjorth parameters with Delta and Beta spectral powers, highlighting the complementarity between the two EEG signal analysis domains.

As future work, it is proposed to validate the optimal subset on other sleep datasets, investigate more sophisticated feature selection methods (SHAP values, RFE), evaluate the impact on alternative classifiers (XGBoost, SVM), and integrate the optimized model into a real-time sleep monitoring PWA, leveraging the computational efficiency of the reduced feature set as a key advantage for embedded deployment.

REFERENCES

GOLDBERGER, A. L. et al. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101, e215–e220, 2000.

GUYON, I.; ELISSEEFF, A. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3, 1157–1182, 2003.

IBER, C.; ANCOLI-ISRAEL, S.; CHESSON, A.; QUAN, S. F. *The AASM Manual for the Scoring of Sleep and Associated Events*. American Academy of Sleep Medicine, Westchester, IL, 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*, 18, 193–202, 2022.

SEKKAL, R. N. et al. Automatic sleep stage classification: From classical machine learning methods to deep learning. *Biomedical Signal Processing and Control*, 77, 103751, 2022.

SNEHA, K. et al. Systematic review: the impact of poor sleep quality on psychological and physical health. *Journal of Heart Valve Disease*, 30, 79–85, 2025.

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*, 7, 513–524, 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*, 12, 286, 2025.