ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

Sleep Staging with Gradient Boosting and DWT-PSD Features from EEG/EOG Signals

Luis Alfredo Moctezuma[1]*, Yoko Suzuki[1], Junya Furuki[1], Marta Molinas[2] and Takashi Abe[1]

1- International Institute for Integrative Sleep Medicine, University of Tsukuba. Tsukuba, Ibaraki, Japan.

2- Department of Engineering Cybernetics, Norwegian University of Science and Technology. Trondheim, Norway.

Abstract. Advances in machine learning (ML) and deep learning (DL) have led to automated sleep staging approaches that achieve high accuracy but often require extensive computational resources and/or high-density electroencephalograms (EEG). This paper presents a method for sleep staging using features extracted via the Discrete Wavelet Transform (DWT) and Power Spectral Density (PSD), followed by the Gradient Boosting (GB) classifier. The study employs a private dataset and the sleep-EDF dataset, comprising EEG and electrooculograms (EOG) channels. The analysis includes configurations with varying numbers of subjects (75, 20, and 12), and the results demonstrate that the proposed method achieves competitive performance with existing approaches that use complex DL architectures, even with fewer subjects. Feature importance analysis highlights the importance of detail coefficients from DWT and PSD-based features from EEG signals. The findings suggest that simplified methods using low-density EEG and EOG with well-selected features and GB classification can offer a viable alternative to DL approaches for sleep staging.

1 Introduction

Sleep staging, which involves identifying wakefulness (W), rapid eye movement (REM) sleep, and three non-REM (NREM) stages (N1, N2, N3), is essential in sleep medicine. Sleep experts manually determine sleep stages based on polysomnography (PSG), which includes signals from electroencephalograms (EEG), electrooculograms (EOG) and electromyograms (EMG). Manual sleep staging is time-consuming and costly, leading to the development of machine and deep learning (ML, DL) models as potential solutions [1–4]. Our previous research has used high-density EEG [3, 4], while other studies have focused on datasets like Sleep-EDF, which include two EEG channels and one EOG channel from over 75 subjects across two sessions [5].

The authors in [6] combined two bipolar EEG channels (Fpz-Cz and Pz-Oz) and one horizontal EOG channel from the Sleep-EDF dataset with EEGNet-BiLSTM, and reported an accuracy of 0.90 and a confusion matrix in which

*This work was partially supported by the Japan Society for the Promotion of Science (JSPS) Postdoctoral Fellowship for Research in Japan: Fellowship ID P22716, JSPS KAKENHI: Grant numbers JP22K19802, JP20K03493, JP22KF0059, and JP22K21351, Japan Agency for Medical Research and Development (AMED) under Grant Number JP21zf0127005, and the Ministry of Education, Culture, Sports, Science and Technology (MEXT), World Premier International Research Center Initiative (WPI) program.

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

the most misclassified classes are N1, N2, and N3. Another recent approach, called SalientSleepNet, was proposed [7] and tested in the SleepEDF-39 and Sleep-EDF-153 datasets using Fpz-Cz EEG and horizontal EOG channels [5], reporting an accuracy of 0.85 on average. Other works used Sleep-EDF-20 (Fpz-Cz channel), Sleep-EDF-78 (Fpz-Cz channel) datasets [5]. In the best case, they obtained an accuracy of 0.856 and 0.829, respectively, for each dataset using an attention-based deep learning architecture (AttnSleep) [8]. A recent work proposed the use of cascaded CNN + LSTM and the use of only one bipolar channel from the Sleep-EDF dataset, showing that it is possible to obtain an accuracy of 0.827 using only Fpz-Cz, and 0.797 using Pz-Oz [9]. In all the mentioned works, there are many factors that confound the comparisons, since different approaches are tested under different conditions, such as using a subset of subjects, splitting into different percentages for testing, training and validation (which are not mentioned in the reported papers), and using a different number of folds in the cross-validation.

Although deep learning methods have shown high performance for sleep staging [3, 7, 8], there is a need to use computationally inexpensive methods that can obtain similar performance while using low-density EEG [10]. Here, we present a method for the classification of 5-class sleep stages by extracting features using the Discrete Wavelet Transform (DWT) and Power Spectral Density (PSD) estimated via Welch's method and then using those features as input to Gradient Boosting (GB) algorithm. Our analysis comprises a set of configurations that incorporate EEG and EOG channels to show the performance obtained with configurations used in the state-of-the-art (using 75 or 20 subjects), as well as using 12 subjects in order to compare the performance in the sleep-EDF dataset and a private dataset.

2 Materials and Methods

2.1 EEG dataset and preprocessing

For our study, we used a private dataset referred to as IIIS-data and a public dataset referred to as sleep-EDF. Relevant information for our analysis is explained below.

The sleep-EDF dataset contains data from 82 healthy subjects for most of them collected during two sessions using 2 EEG channels (Fpz-Cz and Pz-Oz) and 1 EOG channel (EOG-horizontal), which were sampled at 100 Hz. The sleep stages were manually labeled every 30 seconds. To reduce class imbalance, we cut each recording to retain only 30 minutes of wake time prior to the first sleep stage and 30 minutes after the last sleep stage. We only considered subjects with data from the two sessions, thus obtaining 75 subjects after excluding 8 subjects *(Ids: 13,36,39,52,68,69,78,79 as of November 2024)* [5].

IIIS-data was collected at the International Institute for Integrative Sleep Medicine of the University of Tsukuba, Japan [3]. It consists of EEG recordings of 12 subjects (4 females, 22.5± 0.9 years), who slept for ∼8 hours while their data were collected using the BioSemi headcap, a sample rate of 1024 Hz with 128 EEG channels, three EOG channels, three EMG channels, and two mastoid channels. The sleep stages were manually labeled every 30 seconds

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

by a registered polysomnographic technologist. All channels are monopolar; therefore, we applied a process to create bipolar channels using the MNE python library [11]. Using Fpz and Cz channels to obtain Fpz-Cz, Pz and Oz to obtain Pz-Oz, and EOG-l and EOG-r to obtain the EOG-horizontal channel. We also downsampled the data to 100 Hz; in this way, we have the same EEG channels and sample rate in both datasets, making the performance comparison clearer for our experiments.

Experimentally, we defined a bandpass filter from 0.4 to 30 Hz, which was applied to both datasets. For all cases, we used the 30-second segments of the EEG/EOG data for feature extraction.

2.2 Feature extraction, data augmentation and classification

Using DWT with 4 levels of decomposition and the mother wavelet biorthogonal 2.2, EEG and EOG signals were decomposed into approximation coefficients and detail coefficients (*cA4, cD4, cD3, cD2, cD1*). For each subband of each channel, we computed 10 features: Selvik fractal dimension, Katz fractal dimension, Petrosian fractal dimension, Higuchi fractal dimension, instantaneous energy, Teager energy, Hjorth mobility, Hjorth complexity, kurtosis and skewness.

PSD features were calculated to obtain five features per EEG and EOG channel that correspond to delta (0.5-4), theta (4-8), alpha (8-12), sigma (11.5-15.5), and beta (15.5-30).

To address class imbalance in both datasets, we applied data augmentation techniques by tripling the original number of instances for the N1 and REM sleep stages. This approach provided the highest performance; however, various other configurations and sleep stages were also tested. The tested methods for data augmentation were the Synthetic Minority Oversampling Technique (SMOTE), Adaptive Synthetic Sampling (ADASYN), or adding Gaussian noise; however, for comparative purposes and based on the performance obtained, we used ADASYN for both datasets.

Features from EEG/EOG channels were input into various classifiers, validated with a 10-fold cross-validation, ensuring no subject's data appeared in both training and test sets (i.e., 70% of subjects for training, 30% for testing). Model performance was assessed with accuracy, F1score, precision, recall, area under the receiver operating characteristic (AUROC), and kappa. We report only GB results, which achieved the highest performance. Feature importance, calculated as the normalized sum of the criterion's reduction (impurity-based or Gini importance), identified the most relevant channels/features for classification. The higher the value of feature importance, the more critical the feature is for the classification task.

3 Results

3.1 Assessing the performance on sleep-EDF dataset using the 75 subjects

Here, we evaluated the performance using the features from one channel at a time, combining features from one EEG with EOG features, and the two EEG with the EOG channels. In this way, we showed the performance using a high number of subjects with data from the two sessions of the sleep-EDF.

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

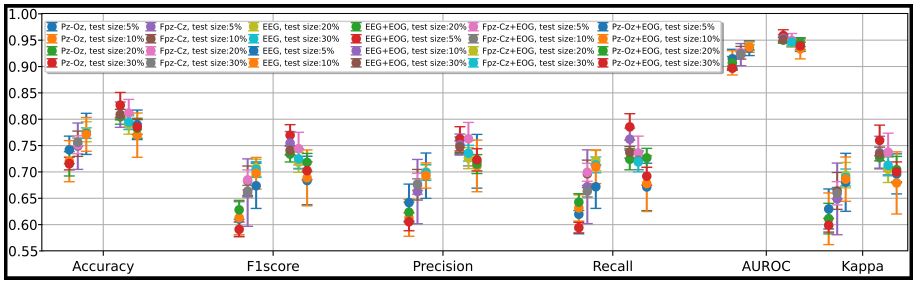

Fig. 1: Performance using all subjects of sleep-EDF using different test set sizes.

Furthermore, we tested performance using different percentage distributions for training and testing of GB models. Fig. 1 shows the results obtained for each metric in the different configurations. The left side of each metric in the plot shows the performance with features of Pz-Oz, Fpz-Cz, or both EEG channels. The right side of each metric in the plot shows the performance using the features of both EEG channels + EOG, Fpz-Cz + EOG, or Pz-Oz + EOG.

The results show that when using a single channel, the highest performance is obtained using the Fpz-Cz, but lower than when using both EEG channels. It also shows that by adding the EOG channel features to any of the EEG channels, performance is best increased to 10% on many metrics. On average in 10-fold, when training the models with data from 95% of the subjects, the performance is similar to that when using data from 70% of the subjects.

3.2 Classification performance in both datasets under different configurations

Based on the previous experiment, this and the following experiments, we decided to use 90% of the data for training and 10% for testing, in all cases we use the two EEG channels and the EOG channel. Here, we present a set of experiments using both datasets with data from 75, 20 and 12 of the subjects from sleep-EDF. In this way, we can compare the results from the state-of-the-art which use 75 or 20 subjects, as well as compare the performance using 12 subjects in both datasets. Table 1 shows the performance obtained in the different metrics and configurations. Other approaches that use the sleep-EDF dataset perform their analysis with 78 subjects from the datasets; however, as we mentioned before, some of the subjects do not have the two recording sessions, which were excluded for our study.

Table 1 shows that when we use data from 75 or 20 subjects, the performance is similar in accuracy and kappa, but around 4% lower for F1score. This shows that even when the models are trained with data from fewer subjects, the performance in the test sets across the 10-fold cross-validation remains similar.

Surprisingly, using data from the first 12 subjects of the sleep-EDF, the accuracy and kappa metrics are higher than even using the 75 subjects, but the F1score remained at 0.77, which may be related to class imbalance in the subjects used, especially for N1 and REM. The point that we want to highlight is that when using the IIIS-data, the performance is lower in most of the metrics, especially for accuracy, F1score, and kappa metrics.

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

Table 1: Classification performance with different number of subjects for sleep-EDF, and 12 subjects of the IIIS-data.

Dataset	Accuracy	F1score	Precision	Recall	AUROC	Kappa
Sleep-EDF, 75 subjects	0.827±0.02	0.770±0.02	0.763±0.02	0.786±0.02	0.959±0.01	0.760±0.03
Sleep-EDF, 20 subjects	0.827±0.01	0.728±0.02	0.736±0.02	0.742±0.01	0.949±0.01	0.750±0.02
Sleep-EDF, 12 subjects	0.844±0.03	0.774±0.04	0.800±0.04	0.776±0.05	0.969±0.01	0.782±0.04
IIIS-data, 12 subjects	0.799±0.05	0.765±0.04	0.774±0.04	0.776±0.04	0.955±0.02	0.731±0.07

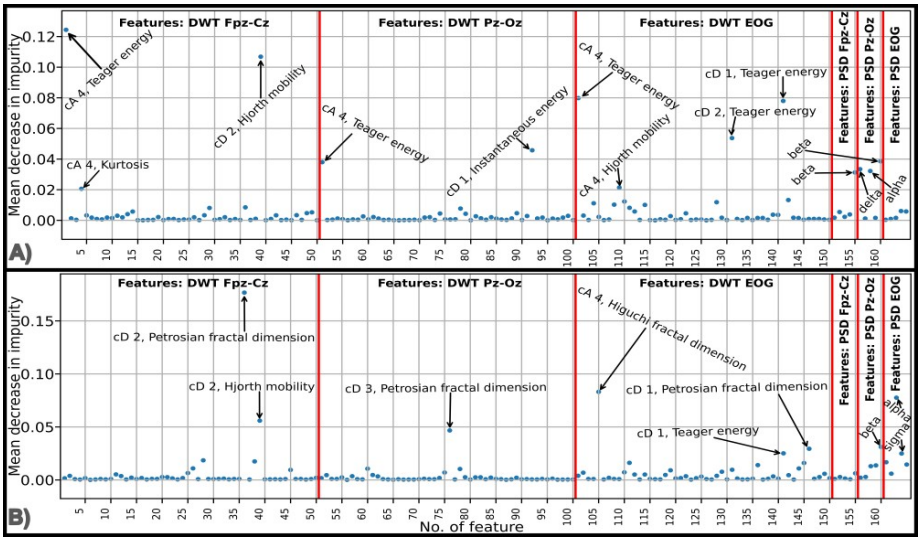

Fig. 2: Impurity-based feature importance for the results presented in Table 1 with 12 subjects in the sleep-EDF dataset (***A)***) and IIIS-data (***B)***).

Fig. 2 shows the impurity-based feature importance for the results obtained with 12 subjects in the sleep-EDF and IIIS-data. It shows the importance of the features on each of the EEG and EOG channels (see Section 2.2), as well as the features based on DWT and PSD.

The Figs. highlight features exceeding a 0.02 threshold, as they are clearly higher than the other features. For the EEG channels, in the case of Fig. 2 *A)*, it shows that the important features are from both the approximation coefficients (low frequencies) and the detail coefficients (high frequencies), illustrated as *cA4, cD1, and cD2*. In contrast, Fig. 2 *B)* indicates important features only from the detail coefficients (illustrated as *cD2 and cD3*). For the case of the PSD-based features obtained from EEG signals, the features are clearly higher for the sleep-EDF dataset. On the other hand, for the EOG features obtained with DWT there are similar features in both datasets, but only for the PSD features from EOG are more important only for the IIIS-data (see Fig. 2 *B)*).

4 Conclusion

Our experiments evaluated the performance of a sleep stage classification method using a combination of DWT and PSD features with the GB algorithm. The proposed method achieves competitive performance with the state-of-the-art approaches using the sleep-EDF, but slightly lower under the same settings for

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

the IIIS-data. The results show that the method is robust to the number of subjects used, with a similar performance obtained using 75, 20, and 12 subjects from the sleep-EDF dataset. The feature importance analysis reveals that the most important features are from the detail coefficients of the DWT (cD3, cD2, cD1) and PSD-based features from the EEG signals, indicating the importance of high-frequency components in sleep stage classification. Our findings suggest that simplified methods using low-density EEG with well-selected features and GB classification can provide a viable alternative to deep learning approaches, potentially making sleep staging more accessible and efficient.

References

[1] Alexander Craik, Yongtian He, and Jose L Contreras-Vidal. Deep learning for electroencephalogram (eeg) classification tasks: a review. *Journal of neural engineering*, 16(3):031001, 2019.

[2] Hojat Ghimatgar, Kamran Kazemi, Mohammad Sadegh Helfroush, and Ardalan Aarabi. An automatic single-channel eeg-based sleep stage scoring method based on hidden markov model. *Journal of neuroscience methods*, 324:108320, 2019.

[3] Luis Alfredo Moctezuma, Yoko Suzuki, Junya Furuki, Marta Molinas, and Takashi Abe. Gru-powered sleep stage classification with permutation-based eeg channel selection. *Scientific Reports*, 14(1):17952, 2024.

[4] Luis Alfredo Moctezuma, Yoko Suzuki, Junya Furuki, Marta Molinas, and Takashi Abe. Enhancing sleep stage classification with 2-class stratification and permutation-based channel selection. In *45th Annual International Conference of the IEEE, Engineering in Medicine and Biology Society*, pages 1–4. IEEE, 2024.

[5] Bob Kemp, Aeilko H Zwinderman, Bert Tuk, Hilbert AC Kamphuisen, and Josefien JL Oberye. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the eeg. *IEEE Transactions on Biomedical Engineering*, 47(9):1185–1194, 2000.

[6] In-Nea Wang, Choel-Hui Lee, Hyun-Ji Kim, Hakseung Kim, and Dong-Joo Kim. An ensemble deep learning approach for sleep stage classification via single-channel eeg and eog. In *2020 International Conference on Information and Communication Technology Convergence (ICTC)*, pages 394–398. IEEE, 2020.

[7] Ziyu Jia, Youfang Lin, Jing Wang, Xuehui Wang, Peiyi Xie, and Yingbin Zhang. Salientsleepnet: Multimodal salient wave detection network for sleep staging. *arXiv preprint arXiv:2105.13864*, 2021.

[8] Emadeldeen Eldele, Zhenghua Chen, Chengyu Liu, Min Wu, Chee-Keong Kwoh, Xiaoli Li, and Cuntai Guan. An attention-based deep learning approach for sleep stage classification with single-channel eeg. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:809–818, 2021.

[9] Yujie Tao, Yun Yang, Po Yang, Fengtao Nan, Yan Zhang, Yulong Rao, and Fei Du. A novel feature relearning method for automatic sleep staging based on single-channel eeg. *Complex & Intelligent Systems*, 9(1):41–50, 2023.

[10] Jeroen Van Der Donckt, Jonas Van Der Donckt, Emiel Deprost, Nicolas Vandenbussche, Michael Rademaker, Gilles Vandewiele, and Sofie Van Hoecke. Do not sleep on traditional machine learning: Simple and interpretable techniques are competitive to deep learning for sleep scoring. *Biomedical Signal Processing and Control*, 81:104429, 2023.

[11] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, et al. Meg and eeg data analysis with mne-python. *Frontiers in Neuroinformatics*, 7:267, 2013.