Feature Importance via Shapley Values in Random Forests for Sleep Apnea and Hypopnea Detection*

Giulia Cisotto[1,*], Shayan Sharifi[2], Shahla Sadeghzadehdarandash[2] and and Leonardo Badia[2]

[1]Dept. of Mathematics, Informatics and Geosciences (University of Trieste), Piazzale Europa, 1, 34127 Trieste, Italy
[2]Dept. of Information Engineering (University of Padova), via G. Gradenigo 6/b, 35131, Padova, Italy

Abstract

Sleep disorders such as apnea and hypopnea have significant health implications, and their accurate identification from biological signals such as polysomnography (PSG) or electrocardiogram (ECG) is essential for effective diagnosis and treatment. We propose a new approach to pinpoint the specific features of these signals that best reveal sleep apnea and hypopnea, through a random forest (RF) algorithm and Shapley value analysis. We validated our approach on the St. Vincent's University Hospital dataset, which includes overnight PSG and ECG signals, from which we extracted time and frequency features, capturing indications sleep apnea and hypopnea. We fed these features into the RF model and evaluated the most influential features in the recognition process, which possibly enables better diagnostic approaches and personalized treatment strategies, combining machine learning with interpretability to advance understanding of sleep disorders.

Keywords

Machine learning, Feature extraction, Random forest classifier, Shapley value, Sleep Apnea, Hypopnea,

1. Introduction

Obstructive sleep apnea syndrome (OSAS) is a very common disorder with an incidence estimated at 5 to 14 percent among adults aged 30 to 70 years. The clinical importance of OSAS is related to an increased risk of cardiovascular disease, as well as higher morbidity and mortality [1]. The gold standard for the diagnosis of OSAS is the polysomnography test (PSG) [2] which provides information on the severity of OSAS and the degree of sleep fragmentation. However, PSG requires an overnight evaluation in a sleep laboratory, dedicated systems, and attending personnel [3].

Recently, medicine has embraced innovative data science techniques, especially those based on machine learning (ML), to effectively analyze vast volumes of clinical data. This aims to deepen our understanding of diseases and enhance diagnostic capabilities.

In this spirit, we employ a supervised ML approach, specifically a random forest (RF) classifier [4], to address the classification of sleep apnea and hypopnea. The dataset is taken from St. Vincent's University Hospital [5]. It contains a ground-truth classification between apnea and hypopnea conditions, in three different categories: obstructive (O), central (C), mixed (M). Thus, the classification is in five classes (HYP-0/C/M, APNEA-O/C), as no samples were collected for the APNEA mixed class. Then, we calculate the Shapley value for each signal to determine their respective contributions to the final diagnosis [6]. This allows us to identify the specific features of each signal that have the greatest impact on sleep apnea and hypopnea. Our Shapley value computations are performed through SHAP (Shapley Additive explanations), a popular software tool for the explainability of ML models [7].

Several studies have explored machine learning for the detection of sleep apnea, highlighting RF as a strong choice for both accuracy and interpretability. Sharaf [8] demonstrated that RF outperforms support vector machines (SVMs) and decision trees in apnea detection based on electrocardiogram (ECG), achieving 91.65% accuracy. This classification task was performed on the Physionet Apnea-ECG

Ital-IA 2025: 5th National Conference on Artificial Intelligence, organized by CINI, June 23-24, 2025, Trieste, Italy
*Corresponding author.
giulia.cisotto@units.it (G. Cisotto); shayan.sharifi@studenti.unipd.it (S. Sharifi); shahla.sadeghzadehdarandash@studenti.unipd.it (S. Sadeghzadehdarandash); leonardo.badia@unipd.it (a. L. Badia)
https://sites.google.com/view/giulia-cisotto/ (G. Cisotto); https://www.dei.unipd.it/~badia/ (a. L. Badia)
0000-0002-9554-9367 (G. Cisotto); 0000-0001-5770-1199 (a. L. Badia)

dataset, sorted into three groups: Apnea (A), Borderline (B), and Normal (C). This study emphasized the role of feature selection, employing sequential feature selection (SFS) and principal component analysis (PCA) to identify the most relevant features. We argue that Shapley values, as we apply in this contribution, would be an even better choice for interpretability based on ML. Bedoya *et al.*[9] further validated RF by comparing it with other ML models, showing that ensemble methods achieve the highest accuracy of around 90%. These results were obtained in a binary classification setting, where the dataset was divided into OSAHS positive (Hypopnea Index > 5) and OSAHS negative (Hypopnea Index < 5). Osa-Sanchez *et al.* [10] reviewed AI-based sleep apnea detection and noted that deep learning models require extensive data and high computational resources, making them impractical for many real-world applications. They also highlighted RF as an efficient alternative that balances accuracy, computational cost, and ease of interpretation.

In general, these studies highlight the effectiveness of RF in detecting sleep apnea through ML for its precision, efficiency, and interpretability. They also show the importance of feature selection in improving model performance.

There are also papers exploring Shapley values and their application to the specific case of sleep apnea. In particular, Tsai *et al.* [11] considered a collection of anthropometric data from a set of Taiwanese patients, with the aim of avoiding time-consuming polysomnography (PSG), whereas Maniaci *et al.* [12] analyzed the importance of clinical scores. In both cases, Shapley values are used for a reduction in dimensionality of features to the most important, enhancing interpretability in research on ML-based sleep apnea.

Troncoso-García et al. [13] trained various classical ML models (logistic regression, RF, etc.) on multi-signal PSG segments for binary apnea event detection, then applied LIME to explain feature contributions. In contrast, our work tackles multi-class apnea-hypopnea classification with multiple signals, moving beyond their single-task, post-hoc RF model approach.

Maniaci et al. [14] developed an SVM-based model to predict OSA severity using only clinical attributes (e.g. BMI, tonsil grade, age, etc.), with SHAP analysis highlighting risk factors like dyslipidemia, palate shape, neck circumference, and BMI as top predictors. Unlike our study, which leverages physiological signal data for event-level detection, their approach relies purely on demographic/clinical features and does not differentiate individual apnea vs. hypopnea events.

Finally, Shi et al. [15] built an interpretable ML ensemble (e.g. XGBoost) to identify severe OSA cases based on questionnaires and anthropometric data, using SHAP plots to rank key features (waist and neck size, ESS score, age, etc.) influencing the model. They focused on binary risk stratification (severe vs. non-severe) using no physiological signals. On the other hand, our approach handles detailed apnea/hypopnea event detection from biosignals, providing finer-grained diagnostic insight.

2. Materials and Methods

2.1. Dataset

St. Vincent's University Hospital Sleep Apnea Database [5] contains 25 full overnight polysomnograms with simultaneous three-channel Holter ECG, from adult subjects with suspected sleep-disordered breathing. PSG is the gold standard test for sleep disorder diagnosis [2] and it records multiple channels, using the Jaeger-Toennies system. Signals recorded were: EEG (C3-A2), EEG (C4-A1), left EOG, right EOG, submental EMG, ECG (modified lead V2), oro-nasal airflow (thermistor), ribcage movements, abdomen movements (uncalibrated strain gauges), oxygen saturation (finger pulse oximeter), snoring (tracheal microphone) and body position. Three-channel Holter ECGs (V5, CC5, V5R) were recorded using a Reynolds Lifecard CF system [5]. The dataset labels, which indicate apnea and hypopnea event types, were pre-assigned by medical professionals in the original St. Vincent's University Hospital Sleep Apnea Database. These annotations were made according to established polysomnography (PSG) guidelines, ensuring accurate and standardized classification of respiratory events. The dataset consists of six predefined event categories—Hypopnea (HYP-O, HYP-C, HYP-M) and Apnea (APNEA-O,

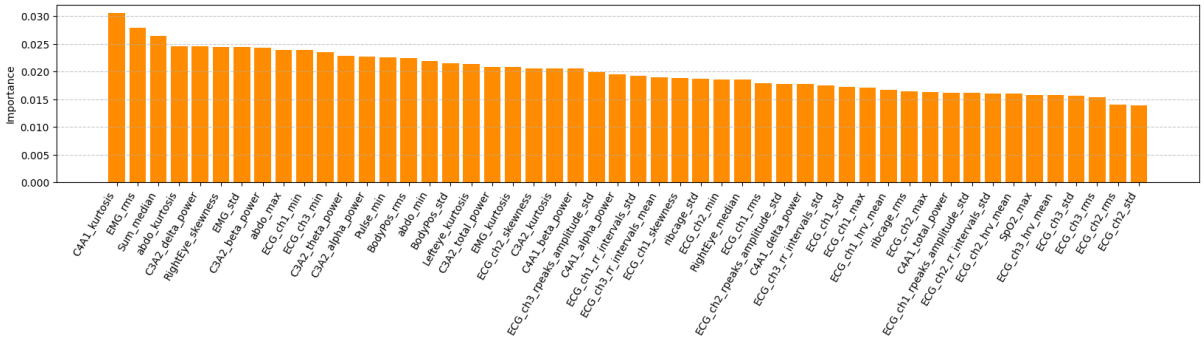

Figure 1: Top 50 features selected and sorted based on their statistical significance (output of SelectKBest Python's method).

APNEA-C, APNEA-M)—which we adopted without any modification. However, the APNEA-M class has no samples, thus we were left with five classes, only.

2.2. Preprocessing and classification

As the dataset was already cleaned from artifacts and noise, we operated a *segmentation* step, only. Signals were tokenized based on pre-annotated respiratory events. Labels, assigned by clinical experts, were available for every segment including an individual respiratory event. Thus, segments can be variable in length, depending on the duration of the corresponding respiratory event. Finally, we applied z-score normalization on each segment to avoid bias toward high-amplitude signals and ensure comparability across channels or modalities.

We extracted a total of 170 features from segments coming from both PSG and Holter ECG signals. Among them, we computed mean and standard deviation from all types of signals, a number of heart rate variability (HRV) related features, oxygen saturation signal (SpO2), EEG-based spectral power in the most common frequency bands (delta, theta, alpha, beta).

To reduce the dimensionality of the dataset, we employed the ANOVA univariate statistical test, as implemented in the *SelectKBest* method included in Scikit-learn open-source Python library [16, 17]. This step allowed us reducing the set of features from 170 to 50. Fig. 1 shows the resulting top-50 features after applying the SelectKBest method: most of them derived from brain activity (EEG), heart activity (ECG), and sensors located at the chest (i.e., EMG_rms, abdo_kurtosis).

The selected features then served as input to the Random Forest (RF) classifier, used to learn patterns and make predictions of the different OSAS levels in the dataset.

Random Forest (RF) is a well-known machine learning classifier, which takes advantage of multiple decision trees, each trained on random permutations of features [6]. Each tree is trained on a different subset of the training set, and the final prediction output is averaged. We randomly split the dataset, assigning 80% of the samples for training and the remaining 20% to the test set. To evaluate the performance of the model, we report the average accuracy across all classes.

2.3. SHAP-based model explaination

Shapley Additive explanations (SHAP) is a model-agnostic interpretability technique used to provide post-hoc explanations for the outcomes of a machine learning model [18, 6]. The Shapley value is a concept from cooperative game theory used to fairly distribute the total gain (or prediction) among individual features based on their contribution to the outcome [19]. In machine learning, it quantifies how much each feature contributes to a model's prediction by computing the average marginal contribution of a feature across all possible feature subsets. Therefore, SHAP is often adopted as feature selection method to identify the most relevant features and/or assess the contribution of each feature to the prediction. Here, we used SHAP to identify the most influential features contributing to the classification of apnea or hypopnea, while accounting for feature interactions.

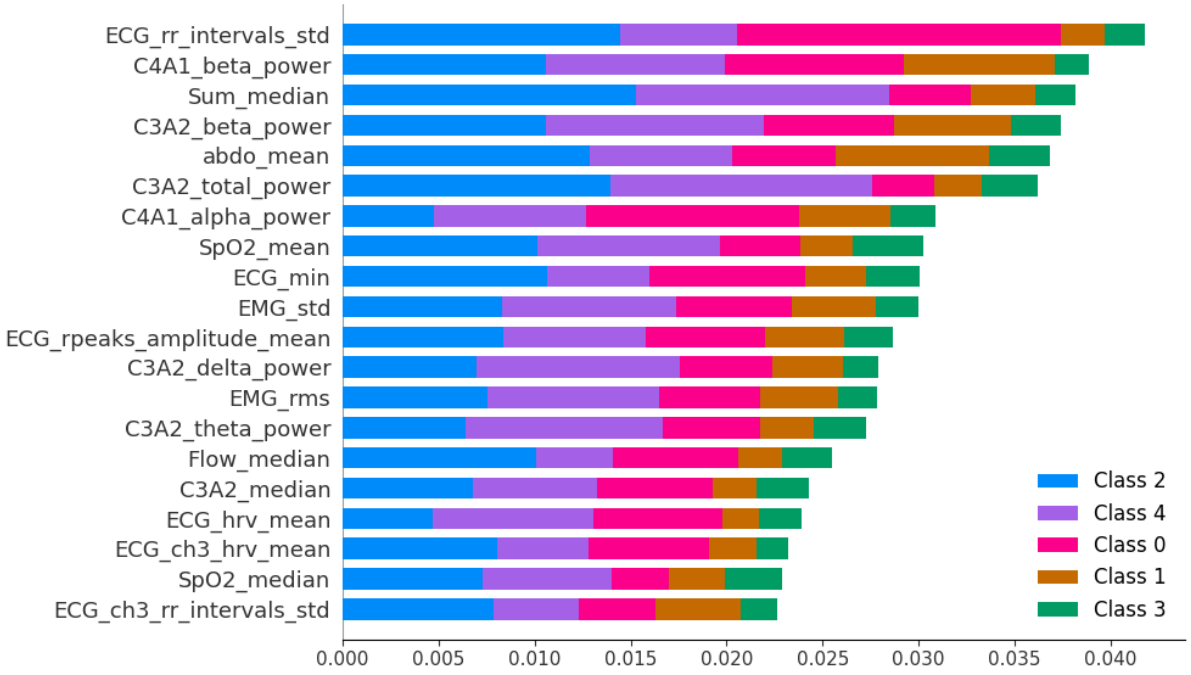

Figure 2: Feature relevance ranking obtained by applying SHAP.

Mathematically, it is calculated by evaluating the change in the model's prediction when a feature is added to every possible subset of other features, and averaging these contributions weighted by the size of the subsets [7]. Therefore, to compute the exact Shapley values, a significant computational effort is required. Although a recent work has introduced an algorithm to compute the exact Shapley values in polynomial time [20], it is convenient to keep the number of input features as small as possible. Thus, we input the top-50 features selected during the ANOVA step.

3. Results and Discussion

We first assessed the performance of our RF classifier. We trained it to classify each data segment into one out of five classes of apnea and hypopnea (APNEA-O/C and HYP-0/C/M, respectively) using 80% dataset and the top-50 features selected using the SelectKBest method. After training, the classifier achieved an accuracy value of approximately 76% (chance level of 20%), indicating its satisfactory ability to correctly classify the five different severity and types of OSAS disease.

Then, we applied SHAP to rank features based on their impact in this specific classification task and this particular classifier. Fig.2 shows the most relevant features as determined by this method.

As expected, the most relevant features for the classification of apnea and hypopnea conditions are the ECG R-R interval standard deviation, which is strictly connected with the heart rate variability (HRV), and several EEG features, mostly related to the beta band ($13-30$ Hz) power from corresponding electrodes in the two hemispheres (C3 and C4). Additionally, the features called *Sum_mean* and *abdo_mean* are identified by SHAP. This is also in line with expectations, since *abdo_mean* stands for *abdominal mean* and reflects the mean amplitude of abdominal respiratory movements over a period, while *Sum_mean* represents the combined respiratory effort of both thoracic (ribcage) and abdominal motion, providing a global measure of breathing effort.

Further contributing features include those extracted from the pulse oximeter that quantifies the saturation level of oxygen in the blood (SpO_2 mean, SpO_2 median), and it is strictly connected with respiration and heart activity. Finally, other features from ECG and EMG complete the set of the most impactful features for the classification.

Comparing feature selection based on ANOVA (Fig. 1) and that obtained via SHAP (Fig. 2), we

can notice a certain degree of agreement. However, the former also includes features related to eye movements and EMG that are expected to be more correlated with disturbed sleep with nocturnal movements, but less with purely respiratory abnormalities. Thus, we can conclude that the set of features identified by SHAP provides a more reliable explanation of the classifier's performance. Future investigations may include a more systematic comparison of subsets of features selected by the two methods to assess performance degradation when removing features that the two methods disagree on.

4. Conclusions

We used Shapley values to explain the classification output from a Random Forest model trained to recognize five different categories of obstructive sleep apnea syndrome. Our analysis showed that certain physiological signals play a crucial role in determining the risk of sleep disorders and have the greatest influence on model predictions. Key contributing factors include heart rate variations (ECG R-R interval), brain activity (C4A1 and C3A2 beta power), and breathing patterns (abdominal mean), among others. Using these features for classification is expected to provide high accuracy and reliable predictions, which make them valuable for future studies and alternative classification methods. Our work stemmed from well-grounded literature that has already suggested Random Forest as a strong candidate for this task. However, it advances the current state of the art by tackling a five-class classification problem and leveraging the full set of physiological signals available in polysomnographic recordings. In future work, we aim to deepen the analysis of feature impact across individual patients and assess the robustness of our findings using different classification models.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] C. Mencar, C. Gallo, M. Mantero, P. Tarsia, G. E. Carpagnano, M. P. Foschino Barbaro, D. Lacedonia, Application of machine learning to predict obstructive sleep apnea syndrome severity, Health Inform. J. 26 (2020) 298–317.

[2] V. K. Kapur, D. H. Auckley, S. Chowdhuri, D. C. Kuhlmann, R. Mehra, K. Ramar, C. G. Harrod, Clinical practice guideline for diagnostic testing for adult obstructive sleep apnea: an american academy of sleep medicine clinical practice guideline, J. Clin. Sleep Med. 13 (2017) 479–504.

[3] B. Pang, S. Doshi, B. Roy, M. Lai, L. Ehlert, et al., Machine learning approach for obstructive sleep apnea screening using brain diffusion tensor imaging, J. Sleep Res. 32 (2023) e13729.

[4] A. T. Azar, H. I. Elshazly, A. E. Hassanien, A. M. Elkorany, A random forest classifier for lymph diseases, Comp. Methods Prog. Biomed. 113 (2014) 465–473.

[5] C. Heneghan, St. Vincent's university hospital/university college Dublin sleep apnea database, 2011.

[6] D. Scapin, G. Cisotto, E. Gindullina, L. Badia, Shapley value as an aid to biomedical machine learning: a heart disease dataset analysis, in: Proc. IEEE Int. Symp. Cluster Cloud Internet Comput. (CCGrid), 2022, pp. 933–939.

[7] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Adv. Neur. Inf. Proc. Syst. 30 (2017).

[8] A. I. Sharaf, Sleep apnea detection using wavelet scattering transformation and random forest classifier, Entropy 25 (2023) 399. doi:10.3390/e25030399.

[9] O. Bedoya, S. Rodríguez, J. P. Muñoz, J. Agudelo, Application of machine learning techniques for the diagnosis of obstructive sleep apnea/hypopnea syndrome, Life 14 (2024) 587. doi:10.3390/life14050587.

[10] A. Osa-Sanchez, J. Ramos-Martinez-de Soria, A. Mendez-Zorrilla, I. Oleagordia Ruiz, B. Garcia-Zapirain, Wearable sensors and artificial intelligence for sleep apnea detection: A systematic review, submitted to J. Health Informat. (2025).

[11] C.-Y. Tsai, H.-T. Huang, H.-C. Cheng, J. Wang, P.-J. Duh, et al., Screening for obstructive sleep apnea risk by using machine learning approaches and anthropometric features, Sensors 22 (2022) 8630.

[12] A. Maniaci, P. M. Riela, G. Iannella, J. R. Lechien, I. La Mantia, et al., Machine learning identification of obstructive sleep apnea severity through the patient clinical features: a retrospective study, Life 13 (2023) 702.

[13] A. d. R. Troncoso-García, M. Martínez Ballesteros, F. Martínez-Álvarez, A. Troncoso, Explainable machine learning for sleep apnea prediction, in: Proc. 26th Int. Conf. on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2022), volume 207, 2022, pp. 2930–2939. doi:10.1016/j.procs.2022.09.351.

[14] A. Maniaci, P. M. Riela, G. Iannella, J. R. Lechien, I. La Mantia, M. De Vincentiis, G. Cammaroto, C. Calvo-Henríquez, M. Di Luca, C. Chiesa Estomba, A. M. Saibene, I. Pollicina, G. Stilo, P. Di Mauro, A. Cannavicci, R. Lugo, G. Magliulo, A. Greco, A. Pace, G. Meccariello, S. Cocuzza, C. Vicini, Machine learning identification of obstructive sleep apnea severity through the patient clinical features: A retrospective study, Life 13 (2023) 702. doi:10.3390/life13030702.

[15] Y. Shi, Y. Zhang, Z. Cao, L. Ma, Y. Yuan, X. Niu, Y. Su, Y. Xie, X. Chen, L. Xing, X. Hei, H. Liu, S. Wu, W. Li, X. Ren, Application and interpretation of machine learning models in predicting the risk of severe obstructive sleep apnea in adults, BMC Medical Informatics and Decision Making 23 (2023). doi:10.1186/s12911-023-02331-z.

[16] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Machine Learning Res. 3 (2003) 1157–1182.

[17] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, et al., Scikit-learn: Machine learning in python, J. Machine Learning Res. 12 (2011) 2825–2830.

[18] S. Ahmed, M. S. Kaiser, M. S. Hossain, K. Andersson, A comparative analysis of LIME and SHAP interpreters with explainable ML-based diabetes predictions, IEEE Access 13 (2024) 37370–37388.

[19] L. S. Shapley, A value for n-person games, Contrib. Th. Games II, Ann. Math. Stud. 28 (1953).

[20] A. Bifet, J. Read, C. Xu, et al., Linear tree shap, Advances in Neural Information Processing Systems 35 (2022) 25818–25828.