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

Giulia Cisotto
Dept. of Mathematics, Informatics and Geosciences
University of Trieste, Italy
giulia.cisotto@units.it

Shayan Sharifi, Shahla Sadeghzadehdarandash, and Leonardo Badia
Dept. of Information Engineering, *University of Padova, Italy*
{shayan.sharifi,shahla.sadeghzadehdarandash}@studenti.unipd.it,
leonardo.badia@unipd.it

***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.**

***Index Terms*—Machine learning; Feature extraction; Random forest classifier; Shapley value, Sleep Apnea; Hypopnea.**

I. 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] that 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 analysis techniques, especially 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 data set is taken from St. Vincent's University Hospital [5]. The dataset 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 six classes (HYP-0/C/M, APNEA-O/C/M).

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 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.

Our approach instead considers a joint analysis of PSG and

This work has been supported by the project D86-RIC-NA-CISOTTO funded by the University of Trieste for newly hired early and mid-career researchers.

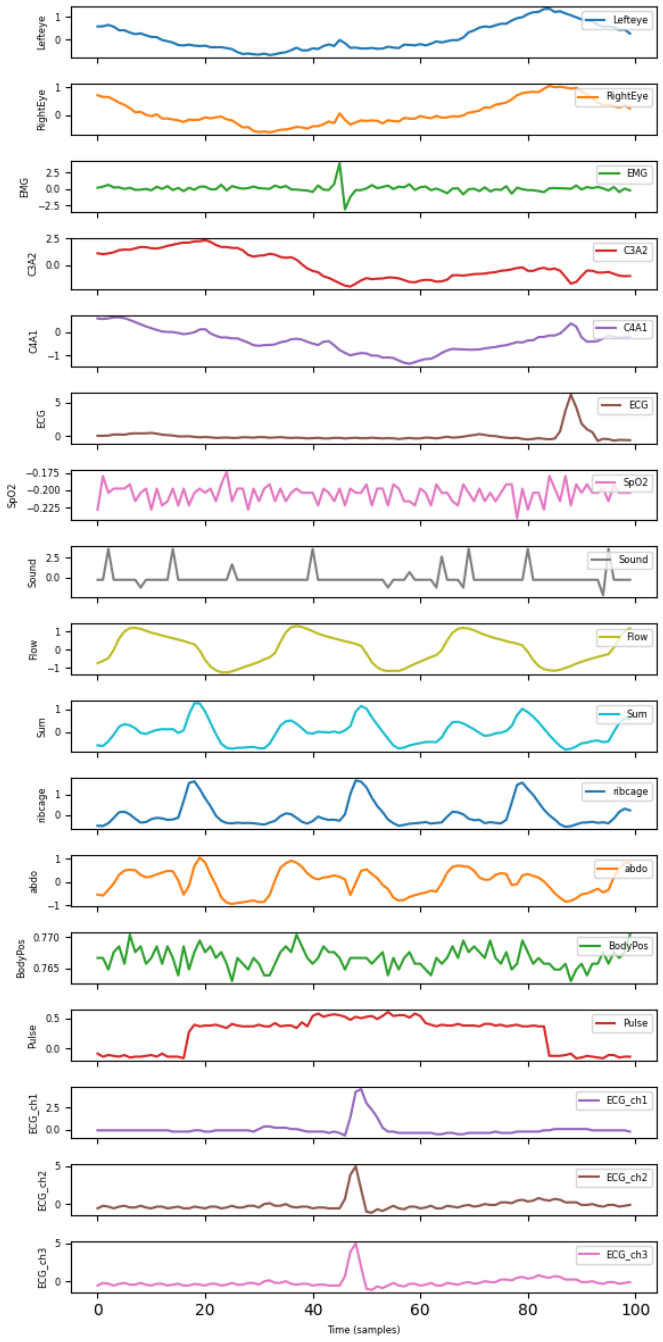

Fig. 1. A sample of few seconds of the signals in the dataset

ECG signals, with a much larger set of available features. Although these can already be compressed through standard techniques of dimensionality reduction, we aim at showing how using Shapley values allow for further improvements, as well as interpreting the results by giving a physiological characterization of importance to certain features, particularly concerning heart rate variations, brain activity in certain intervals, and overall breathing patterns.

II. Materials and Methods

A. 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]. A sample of the signals in the dataset is shown in Fig. 1.

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, APNEA-C, APNEA-M)—which were directly incorporated into our study without modification.

A RF classifier was applied to this dataset after training it with these labels taken as ground truth. However, the focus of our approach was to analyze the pre-processing, where the signal is segmented, then relevant features are extracted. In particular, our pipeline consists of the following steps.

B. Preprocessing

After PSG and ECG signals are read from the dataset, no additional noise removal or artifact rejection was needed, as the signals of the dataset are pre-cleaned. Thus, we just needed to perform *segmentation*, dividing the signals based on pre-annotated respiratory events. We based it on the available labels assigned by experts, each segment including an individual respiratory event. Thus, segments can be variable in length, depending on the duration of the corresponding respiratory event. Z-score normalization was applied to each physical quantity to ensure consistency across the data, translating and down-scaling each segment to have zero mean and unit variance for all components. We remark that the normalization was computed per segment, not globally.

We extracted relevant *features* from the segmented signals.

In this study, we extracted a total of 170 features from both PSG and Holter ECG signals. Several well-known features are extracted from input signals such as RR interval signals, ECG-derived respiration (EDR) signals, heart rate variability (HRV), oxygen saturation signal (SpO2), blood gas or blood oxygen saturation (SaO2), and autocorrelation function (ACF). From PSG, we also extracted a larger set of time-domain features (e.g., mean, standard deviation, variance, kurtosis)

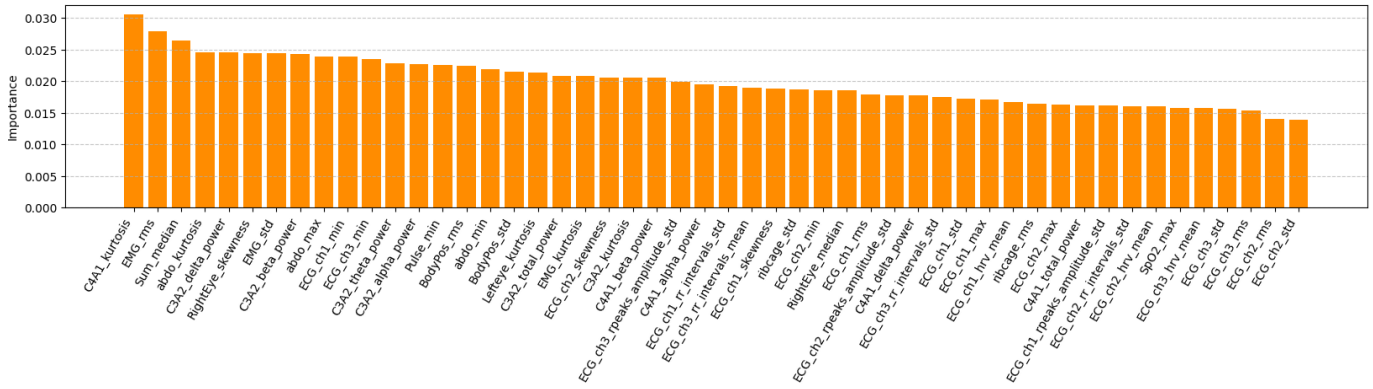

Fig. 2. Top 50 features selected and sorted based on their statistical significance (output of SelectKBest Python's method).

and frequency-domain features (e.g., alpha power, beta power, delta power) computed using Wavelet Transform (WT) [13]. Additionally, we included respiratory-related features, such as SpO_2 mean and respiratory flow variability. From Holter ECG, we extracted a smaller subset of features, mainly focusing on HRV-related metrics such as: (i) difference in the root mean square of RR peak amplitude; (ii) time interval between consecutive HRV signals not exceeding 50 milliseconds; (iii) standard deviation of HRV signals; (iv) mean, variance of ECG signals.

C. Dimensionality reduction

To reduce the dimensionality of the dataset, we employed SelectKBest with ANOVA F-score ($f\_classif$), as implemented in the Scikit-learn open-source Python library [14], selecting the 50 most relevant features. This method evaluates the best features [15] based on the statistical significance in distinguishing between apnea/hypopnea event classes. This helps to eliminate irrelevant or redundant features, improving the efficiency and effectiveness of our classification model. Based on an empirical evaluation, we observed that the best 50 features provided an optimal balance between model performance and computational efficiency. Fig. 2 illustrates the resulting features after applying the SelectKBest method. As expected, the majority of the top-50 informative features are derived from the brain activity (EEG) and the heart activity (ECG), with a remainder contribution from other sensors (EMG for muscular activity and eye movements, and body posture).

Pairwise correlation was applied on the remaining 50 top features to quantify the (linear) relationship between them. For the computation, we used Pearson's correlation $r_{X,Y}$ computed as

$$r_{X,Y} = \frac{\sum_{n=1}^{N}(X(n)-\bar{X})(Y(n)-\bar{Y})}{\sqrt{\sum_{n=1}^{N}(X(n)-\bar{X})^2}\sqrt{\sum_{n=1}^{N}(Y(n)-\bar{Y})^2}}. \quad (1)$$

where X and Y are the two features considered in the pair, while the summation is performed across the total number of segments N. $\bar{X}$ represents the average value of the first feature across all segments, and $\bar{Y}$ similarly for the second feature.

The aim of this step was to evaluate the degree of correlation still present in the dataset and support the need for a more effective feature selection method. Also, visualizing the correlation matrix, we could check for data quality, as we expect higher correlation from sensors capturing the same physiological parameter.

Afterwards, we utilized an RF algorithm for classification. The selected features from the previous step serve as input to the random forest classifier, allowing it to learn patterns and make predictions based on the training data. Finally, we calculated the Shapley values for the trained RF model, which quantify the contribution of each feature towards the final prediction, to gain insight into the importance of different features in the classification.

D. Classifier

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.

To ensure robust classification, we randomly divided the dataset, allocating 80% of the samples for the training and the remaining 20% for validation. For reproducibility, we set the random-state parameter to 42.

E. Classification performance

The classification task in this study is a multi-class problem, distinguishing between six different apnea/hypopnea event types. To evaluate the performance of the classifier, we use standard classification metrics, including *accuracy, precision, recall (sensitivity), F1-score, and specificity*. Given the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) for each class $i = 1, 2, ..., 6$ for each class, we calculated several well-known classification metrics. The accuracy is computed as:

$$\text{Accuracy} = \frac{\sum_i TP_i}{\sum_i (TP_i + FP_i + FN_i)}[\%]. \quad (2)$$

and it represents the overall correctness of the classifier.

Precision measures the number of predicted positive cases that were actually correct for each class:

$$\text{Precision}_i = \frac{TP_i}{TP_i + FP_i}[\%]. \quad (3)$$

Recall (or sensitivity) measures how many actual positive cases were correctly identified:

$$\text{Recall}_i = \frac{TP_i}{TP_i + FN_i}[\%]. \quad (4)$$

Specificity, which measures how well the classifier identifies negative cases, was computed for each class as:

$$\text{Specificity}_i = 100 \times \frac{TN_i}{TN_i + FP_i}[\%]. \quad (5)$$

Finally, the F1-score is the harmonic mean of precision and recall:

$$\text{F1}_i = 100 \times \frac{2 \times \text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i}[\%]. \quad (6)$$

All the above metrics refer to a single class. However, since this is a multi-class classification problem, aggregated metrics can be reported, too. Particularly, macro-averaged metrics are obtained by computing the unweighted mean of the metric across all classes. On the other hand, weighted-averaged metrics are computed by considering class imbalances, also.

Our classification results include precision, recall, F1-score, and specificity per class, as well as macro and weighted averages.

F. Feature selection via Shapley values

Shapley values, originating from cooperative game theory [16], are widely used as a practical quantification instruments by many studies requiring an interpretation of ML results [6], [17]. In particular, they are often adopted in feature selection, to identify the most relevant features and/or assess the contribution of each feature to the prediction. They quantify the marginal contribution of a feature by considering all possible feature subsets, ensuring an explainable and robust feature ranking.

In our problem, we used Shapley values to identify the most influential features contributing to the classification of apnea or hypopnea, while accounting for feature interactions. We used SHAP, a practical package in Python that calculates Shapley values for different ML models [7], including RF. As a result, we identified those features (among the top-50 set) which mostly impact on the classification, providing an explanation on the output of the RF classifier.

III. Results and Discussion

We first assess the performance of our RF classifier. We trained it to classify each data segment into one out of six classes of apnea and hypopnea (APNEA-O/C/M and HYP-O/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%, indicating its satisfactory ability to correctly classify the six classes.

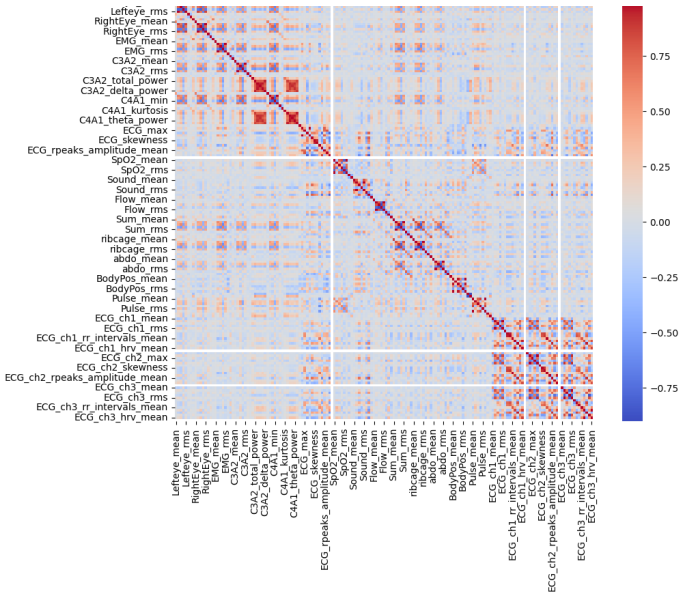

Fig. 3. Correlation matrix including pair-wise Pearson's correlation coefficients computed on the features used for the classification.

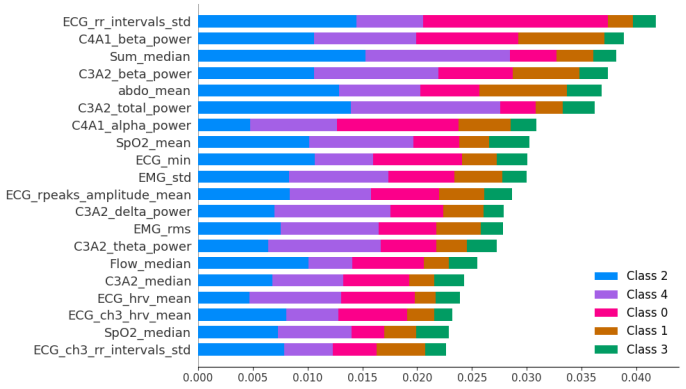

Fig. 4. Feature relevance ranking obtained by applying SHAP.

We then studied the correlation among the 50 features: Fig. 3 shows the correlation matrix. Features derived from the same sensor (e.g., ECG, EEG) tend to be correlated, reflecting the expected physiological relationships. For example, the right bottom corner of the correlation matrix displays particularly correlated values, reflecting the common derivation of those three ECG signals (from the Holter device). However, the correlation between ECG features from the PSG device and the ECG features from the Holter are not equally large. In other works, these features may have been discarded as redundant, but they are actually not. Therefore, we need to deepen our investigation using a more effective feature selection method to identify those sensors which can better capture the pathology-related heart activity pattern.

Therefore, we applied SHAP to rank features based on their impact of our specific classification problem. Fig.4 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.

Additional contributing features include those extracted from the pulse oximeter and quantify the saturation level of oxygen in the blood (SpO_2 mean, SpO_2 median), strictly connected with respiration and heart activity. levels, heart signals, and muscle activity: Blood oxygen levels (): This represents the average oxygen saturation in the blood. 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. 2) and that obtained via SHAP (Fig. 4), 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.

IV. Conclusions

We used Shapley values to explain how our sleep disorder classification model makes its decisions across six categories. 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.

References

[1] C. Mencar, C. Gallo, M. Mantero, P. Tarsia, G. E. Carpagnano, M. P. Foschino Barbaro, and D. Lacedonia, "Application of machine learning to predict obstructive sleep apnea syndrome severity," *Health Inform. J.*, vol. 26, no. 1, pp. 298–317, 2020.

[2] V. K. Kapur, D. H. Auckley, S. Chowdhuri, D. C. Kuhlmann, R. Mehra, K. Ramar, and 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.*, vol. 13, no. 3, pp. 479–504, 2017.

[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.*, vol. 32, no. 1, p. e13729, 2023.

[4] A. T. Azar, H. I. Elshazly, A. E. Hassanien, and A. M. Elkorany, "A random forest classifier for lymph diseases," *Comp. Methods Prog. Biomed.*, vol. 113, no. 2, pp. 465–473, 2014.

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

[6] D. Scapin, G. Cisotto, E. Gindullina, and 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 and S.-I. Lee, "A unified approach to interpreting model predictions," *Adv. Neur. Inf. Proc. Syst.*, vol. 30, 2017.

[8] A. I. Sharaf, "Sleep apnea detection using wavelet scattering transformation and random forest classifier," *Entropy*, vol. 25, no. 3, p. 399, 2023.

[9] O. Bedoya, S. Rodríguez, J. P. Muñoz, and J. Agudelo, "Application of machine learning techniques for the diagnosis of obstructive sleep apnea/hypopnea syndrome," *Life*, vol. 14, no. 5, p. 587, 2024.

[10] A. Osa-Sanchez, J. Ramos-Martinez-de Soria, A. Mendez-Zorrilla, I. Oleagordia Ruiz, and 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*, vol. 22, no. 22, p. 8630, 2022.

[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*, vol. 13, no. 3, p. 702, 2023.

[13] E. S. JeyaJothi, J. Anitha, S. Rani, and B. Tiwari, "[retracted] a comprehensive review: Computational models for obstructive sleep apnea detection in biomedical applications," *BioMed Res. Int.*, vol. 2022, no. 1, p. 7242667, 2022.

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

[15] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Machine Learning Res.*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[16] L. S. Shapley, "A value for n-person games," *Contrib. Th. Games II, Ann. Math. Stud.*, vol. 28, 1953.

[17] S. Ahmed, M. S. Kaiser, M. S. Hossain, and K. Andersson, "A comparative analysis of LIME and SHAP interpreters with explainable ML-based diabetes predictions," *IEEE Access*, vol. 13, pp. 37 370–37 388, 2024.