Explainable Artificial Intelligence in the Estimation of Sleep Apnea Severity Using Heart Rate Variability

Gabriel B. Vitorino*, Rafael R. Santos**, Luiz E. V. Silva***, Alan L. Eckeli**, Rubens Fazan-Junior**, Renato Tinós*

*FFCLRP, University of São Paulo (USP), Ribeirão Preto - SP, Brazil

**FMRP, University of São Paulo (USP), Ribeirão Preto - SP, Brazil

***Children's Hospital of Philadelphia, Philadelphia - PA, United States

gabriel.branco@usp.br, rtinos@ffclrp.usp.br

***Abstract*—An Explainable Artificial Intelligence algorithm is used to generate explanations of the decisions of Machine Learning models trained to classify the severity of Obstructive Sleep Apnea based on Heart Rate Variability. First, a Random Forest classifier is used for Obstructive Sleep Apnea severity prediction and then the Local Rule-based Explanations method is applied to provide decisions interpretability. The method employs Genetic Algorithms to generate artificial examples, which are then used to train decision trees that approximate the local behavior of the black-box model. This approach enables the extraction of human-readable rules that explain individual predictions, offering valuable insights into the decision-making process of black-box models such as Random Forests. By identifying the most influential features and decision rules, our approach aims to aid medical professionals in understanding how Heart Rate Variability correlates with Obstructive Sleep Apnea severity and enhances trust in Machine Learning-driven diagnostic tools.**

***Index Terms*—Explainable Artificial Intelligence, Genetic Algorithms, Heart Rate Variability, Obstructive Sleep Apnea**

I. Introduction

Some Machine Learning (ML) algorithms, like Artificial Neural Networks and Random Forests (RFs), have been highly effective in addressing a wide range of complex problems. However, despite their effectiveness, the decisions made by these ML algorithms are often difficult for humans to interpret, even for domain experts. It is worth noting that, in this context, interpretability is defined as the ability to explain or provide meaning to the decisions made by ML models in terms that are understandable to humans, resembling our own decision-making process [1].

This is not a simple task, as many models behave like black-boxes, that is, models whose decisions are not easily explainable, either because their internal mechanisms are obscure, such as in proprietary software, or because their decisions are inherently difficult for humans to interpret, which is the case of Deep Learning algorithms. Regarding the decisions made by such models, practically all that is known is the mapping between input and output data. On the other hand, interpretable ML models, such as Decision Trees (DTs), generate interpretable decisions when inputs are transformed into outputs. In many fields, such as Medicine, sacrificing interpretability for higher accuracy is not ideal, as it is essential that the predictive model is both accurate and transparent.

However, there is an inherent trade-off between performance and interpretability, as the best-performing models are often the least interpretable, while the most interpretable models tend to have lower performance. Furthermore, with the growing adoption of Artificial Intelligence (AI), new regulations have been introduced to govern its use, such as the brazilian *Lei Geral de Proteção de Dados Pessoais* (LGDP) [2]. Article 20 of the LGPD defines the right to explanation and ensures that any individual affected by an automated algorithmic decision has the right to request reconsideration and understand why that decision was made.

Recent studies have sought to address this issue by proposing techniques that explain the decisions made by

This work was partially supported by São Paulo Research Foundation - FAPESP (under grants #2024/04927-6, #2024/15430-5 and #2024/08485-8) and National Council for Scientific and Technological Development - CNPq (under grant #304640/2024-7).

black-box models in an interpretable manner [3]. These techniques belong to the research area known as Explainable AI (XAI). There are a variety of XAI methods, each with different characteristics, advantages, and limitations [4]. One of the most popular XAI methods is Local Interpretable Model-agnostic Explanations (LIME) [5]. LIME is a local explanation method that assumes the decision boundaries of ML models are locally linear. Therefore, LIME fits a simple linear model in the neighborhood of a single instance to be explained by a black-box ML model. The linear model locally approximates the behavior of the more complex ML model and can thus be used to provide a local explanation for its prediction. LIME is also agnostic, what means that it can explain the decisions of any ML model.

In this work, we propose to use the Local Rule-based Explanations (LORE) method to provide interpretability to the classification of Obstructive Sleep Apnea (Section II) made by black-box models. Similarly to LIME, LORE (Section III) is a local explanation and agnostic method. LORE uses Genetic Algorithms to generate artificial examples, which are then used to train decision trees that approximate the local behavior of the black-box model. This approach enables the extraction from a DT of human-readable rules that explain individual predictions. The methodology, experimental results and conclusions are respectively presented in sections IV, V, and VI.

II. Obstructive Sleep Apnea

Obstructive Sleep Apnea (OSA) is the most common sleep disorder, characterized by repetitive events of partial or complete obstruction of the upper airways during sleep [6]. These obstructions result in recurrent episodes of hypoxia and hypercapnia, causing significant physiological alterations. The gold standard for diagnosing OSA is polysomnography (PSG), an exam that simultaneously records various physiological signals during sleep, allowing analysis of sleep stages and disorders. The severity of OSA is quantified by the apnea-hypopnea index (AHI), calculated by specialists based on PSG recordings. However, PSG is time-consuming and costly, leading to long waiting lists and under-diagnosis of the disease.

OSA is associated with the development of various comorbidities, especially cardiovascular diseases, and negatively impacts patients' quality of life. Additionally, the sleepiness and fatigue caused by poor sleep quality increase the risk of accidents at work and in traffic, endangering both the patient and others. In this context, new diagnostic techniques are needed to facilitate screening and monitoring of the disease. The use of heart rate variability (HRV) data extracted from electrocardiogram (ECG) recordings has emerged as a promising tool for diagnosing OSA [6], being a noninvasive method capable of assessing cardiac autonomic modulation from time series extracted from the ECG.

III. Explainability Algorithm

LORE [7] aims to explain the decision of any ML model, including black-box models, for a specific instance through a surrogate interpretable model (DT). This is done by locally reproducing the decision boundaries of the black-box model under analysis. The DT is trained using a synthetic dataset generated by Genetic Algorithms (GAs). The DT subsequently provides a set of logical rules and counterfactual rules that explain the decision made by the black-box model for the specific instance.

Two GAs are used: one is responsible for generating artificial instances similar to the one being explained, i.e., that are classified in the same class by the black-box algorithm, while the other is responsible for generating artificial instances that belong to a different class but are still in the surroundings of the instance. In this context, similarity is evaluated using a distance function computed between the instance being explained and the artificial instances.

1: **inputs:**
- $\mathbf{x}$ - instance to explain
- b - black-box model
- N - population size
- G - number of generations
- p_c - crossover rate
- p_m - mutation rate

2: **output:**
- $e_c(\mathbf{x})$ - explanation for decision $y_b(\mathbf{x})$

3: $Z_1 \leftarrow GA_1(\mathbf{x}, b, N/2, G, p_c, p_m)$;
4: $Z_2 \leftarrow GA_2(\mathbf{x}, b, N/2, G, p_c, p_m)$;
5: $Z \leftarrow Z_1 \cup Z_2$;
6: $c \leftarrow DecisionTree(Z)$;
7: $r_c(\mathbf{x}) \leftarrow ExtractRule(c, \mathbf{x})$;
8: $\Phi_c(\mathbf{x}) \leftarrow ExtractCounterfactuals(c, r_c(\mathbf{x}), \mathbf{x})$;
9: $e_c(\mathbf{x}) = \langle r_c(\mathbf{x}), \Phi_c(\mathbf{x}) \rangle$;

Algorithm 1: LORE [7]

The two GAs differ only in the fitness functions, respectively given by:

$$f_1(\mathbf{z}) = \mathbb{I}_{y_b(\mathbf{x})=y_b(\mathbf{z})} + (1 - d(\mathbf{x}, \mathbf{z})) - \mathbb{I}_{\mathbf{x}=\mathbf{z}} \quad (1)$$

$$f_2(\mathbf{z}) = \mathbb{I}_{y_b(\mathbf{x})\neq y_b(\mathbf{z})} + (1 - d(\mathbf{x}, \mathbf{z})) - \mathbb{I}_{\mathbf{x}=\mathbf{z}} \quad (2)$$

where $y_b(\mathbf{x})$ and $y_b(\mathbf{z})$ are respectively the classes assigned by the black-box algorithm to the instance being explained ($\mathbf{x}$) and the artificial instance represented by individual $\mathbf{z}$ of the GA, $d(\mathbf{x}, \mathbf{z})$ is the distance function; and $\mathbb{I}(\text{true}) = 1$ and $\mathbb{I}(\text{false}) = 0$. In the version of LORE used in this work, the similarity term $1 - d(\mathbf{x}, \mathbf{z})$ is also compared to a constant η, which aims to increase

the genetic variability of the data. If the similarity is greater than this constant, the data points are considered too close, and this term is zeroed in the fitness function, thus favoring solutions that better fill the feature space.

In LORE, the DT c, trained by the artificial dataset Z, extracts classification rules, $r_C(\mathbf{x})$, by following the path taken through the tree nodes to the leaf that classifies the instance to be explained $\mathbf{x}$. The algorithm also extracts counterfactual rules $\Phi_c(\mathbf{x})$, which indicate what changes to the rules would lead to a different classification. This is done through a process that aims to minimize the number of changes required to alter the class of the instance to be explained [7]. The pseudocode of LORE is shown in Algorithm 1.

IV. Methodology

In previous studies we have used machine learning models such as Random Forests (RF) and Multilayer Perceptrons (MLP) to effectively classify OSA severity based on HRV, oxygen saturation, and anthropometric data [6], but our focus in this paper is not merely on predictive performance instead, we aim to interpret the decision-making process of the best-performing model—Random Forest using LORE, in order to increase knowledge and trust in the models.

A. Heart Rate Variability Data

The dataset used in this work contains information from 291 polysomnography exams performed between 2015 and 2022 at the *Hospital das Clínicas* affiliated with the University of São Paulo at Ribeirão Preto. The dataset was collected in the course of one of the authors' PhD research, that used ML models to classify OSA [6]. The study was authorized by the Research Ethics Committee for Human Subjects of the same hospital (Protocol: 42058720.6.000.5440/4.550.2327).

The PSG exam records various patient information, including electroencephalogram, electromyogram, electrocardiogram (ECG), pulse oximetry, and nasal pressure. The data obtained from the ECG are used to calculate HRV measures, that will be used to train the black-box models. For each patient, six 15-minute segments are selected, corresponding to the first six hours of the ECG recording. From these segments, the RR intervals (RRi) are calculated, forming time series for each segment. Subsequently, preprocessing and postprocessing steps are applied to these time series. From the RRi series, 34 different HRV indices are extracted using linear and nonlinear methods [6].

Among them are time-domain indices such as the mean of RRi, Standard Deviation of NN intervals (SDNN), and Root Mean Square of Successive Differences (RMSSD). Frequency-domain indices are obtained through spectral analysis with interpolation, windowing, and Fourier transform, yielding power measures in the VLF (Very Low Frequencies; $<$0.04 Hz), LF (Low Frequencies; 0.04–0.15 Hz), and HF (High Frequencies; 0.15–0.4 Hz) bands, as well as the LF/HF ratio, both in absolute and normalized values (LFnu and HFnu).

Regarding nonlinear methods, detrended fluctuation analysis (DFA) is used to capture fractal properties of the RRi series. Seven entropy measures are also calculated—Sample Entropy (SampEn), Fuzzy Entropy (FuzzyEn), Distribution Entropy (DistEn), Attention Entropy (AttEn), Dispersion Entropy (DispEn), Phase Entropy (PhaseEn), and Permutation Entropy (PermEn)—each reflecting different aspects of irregularity and complexity in the time series. Additionally, two symbolic dynamics approaches are applied: Max-Min and binary, both transforming the RRi series into symbolic sequences for autonomic variation pattern analysis. Heart rate fragmentation (HRF) is assessed by identifying inflection points between acceleration and deceleration phases, quantifying transition patterns. Finally, HRV asymmetry indices such as Porta, Guzik, and Ehlers indices are used to characterize differences in acceleration and deceleration dynamics. The Acceleration Capacity (AC) and Deceleration Capacity (DC) indices are also employed to explore autonomic responsiveness of the heart under various conditions [6].

The PSG reports, analyzed by a sleep medicine specialist, provide important information for characterizing OSA and the clinical profile of patients. Three main scores were provided by the specialist: apnea-hypopnea index (AHI), which is used to define the patients' class labels used here, minimum oxygen saturation during sleep (SatMin), and the percentage of total sleep time with oxygen saturation below 90% (T90). SatMin and T90 were also considered as inputs of the black-box models. The AHI is calculated as the average number of apnea and hypopnea events per hour of sleep, with events defined by significant airflow reductions accompanied by oxygen desaturation or arousals. This index is used to define 4 classes: patients with AHI between 5 and 15 were classified as Mild OSA, between 15 and 30 as Moderate, and above 30 as Severe; values below 5 indicated absence of the condition. In addition to these data, anthropometric information including sex, age, height, weight, and body mass index (BMI) was collected to build a more complete clinical profile and enable correlational analyses with HRV data [6]. These anthropometric information was also included as inputed of the black-box models.

B. Model Training

The version of LORE implemented in the XAI-Library [8] is used to generate explanations for the classification of OSA severity levels in different patients, as performed

by an RF model using heart rate variability data, blood oxygen saturation, and other anthropometric variables. The training of the black-box model (RF) is performed using all features listed in previously and is done in the same way proposed in [6]. Some individuals had missing data, which were either estimated using the mean value of that attribute for a given class or calculated, as in the case of BMI, which is determined from the patient's weight and height. The RF was implemented by using the standard version implemented in the SciKit Learn library. More specific details about the different models are provided in the results section. In all cases, data balancing was performed using the Synthetic minority over-sampling technique (SMOTE) to ensure an equal number of individuals in each class, especially because the Severe class initially had many more examples than the others.

TABLE I
ANTHROPOMETRIC AND PSG CHARACTERISTICS BY OSA SEVERITY CLASS. VALUES ARE N (%) FOR MEN, AND MEDIAN FOR OTHER VARIABLES [6].

	Non-Apneic (n=47)	**Mild (n=63)**	**Moderate (n=70)**	**Severe (n=111)**
Men	12 (25.5%)	27 (42.9%)	30 (42.9%)	49 (44.1%)
Age (years)	41	52	56.5	56
Height (m)	1.62	1.66	1.65	1.63
Weight (kg)	71.5	82	85.5	98
BMI	26.7	31.04	31.01	34.72
AHI	2.6	10.0	21.1	57.7
T90	0.07	1.10	2.75	19.80
SatMin (%)	88	86	83	74

The distribution of examples (patients) in the dataset is shown in Table 1. The Non-Apneic class has the smallest amount of individuals with only 47, the next two, Mild and Moderate, are slightly more representative with 63 and 70 individuals, while the Severe class have 111 patients representing a big percentage of training data, which is a problem solved by balancing the classes using SMOTE to maintain a decent amount of training data. After balancing, all classes have 111 individuals, matching the initial number in the Severe class. In all cases, explanations and accuracy metrics are calculated only over the the examples of the original dataset and not over synthetic data.

V. RESULTS

Results of two experiments are presented: one for four-class classification using separate training and testing sets, and; another for a four-class classification model trained using all available examples. Explanations were generated and analyzed for each model.

A. Experiment 1: Training and Test Sets

First, the black-box model (RF) was trained. The original dataset contains 291 samples; and after the described balancing, the number of examples was 444 individuals (111 per class). The dataset was then divided into training and test sets, with 380 instances for training and 64 for testing. This division yielded an accuracy of 68%, calculated over the original examples only and weighted across all classes, a value consistent with previous findings [6]. From the 64 test instances, 45 explanations were generated for the non-synthetic instances.

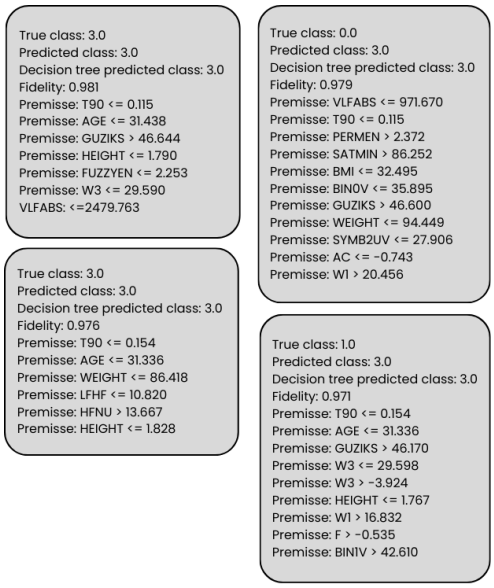

Fig. 1. The first block on the left shows the explanation generated for a Non-Apneic (3) individual correctly classified. The same applies to the second explanation block on the left. In both cases, we observe which premises were met by the examples that led to this classification. On the right the first block shows the explanation generated for a Severe-class (0) individual that was misclassified as Non-Apneic (3), while the second block displays a Non-Apneic (3) prediction for an individual whose true class is Mild (1). In both cases, we observe which premises were satisfied by the samples that led to these classifications.

The analyses were divided between cases where the RF (black-box model) correctly predicted the class of the example and cases where it misclassified the example. The previously defined classes were automatically encoded by the algorithm, with the number 3 representing the Non-Apneic class, 2 representing the Moderate class, 1 representing the Mild class, and 0 representing the Severe class. Figure 1 shows the LORE explanations for four examples. In the figure, we can observe that the predicted classes from the black-box and surrogate models coincide. Each explanation offers a fidelity measure between the models and, the premises satisfied by the example to be explained that led to that particular classification. As we can observe, the T90 parameter having a small value is important for classifying individuals into class 3 (Normal) - being the first premise tested by the surrogate model in three of

the examples. This makes considerable sense since this parameter indicates the percentage of total sleep time during which the patient's oxygen saturation remains below 90% [6].

On the first explanation on the right side of Figure 1 the algorithm commits an error by classifying a Severe patient as Normal, and in the other one it makes another error by assigning the Normal class to a patient with Mild OSA. In the first case, the algorithm initially analyzes VLFABS - a parameter of great importance for this model [6] - but then evaluates T90 using a threshold very similar (if not identical) to the ones used in the previous case to correctly classify the two instances correctly as Non-Apneic, which may be related to it mistakenly classifying the other two on the right as Non-Apneic, even tho they belong to a different class. The second block shows the same T90 threshold analysis also being used to incorrectly classify the individual as Normal. Based on these examples, we can infer that the algorithm has identified a relationship between a low T90 threshold and the Normal class, which leads to both correct classifications and errors.

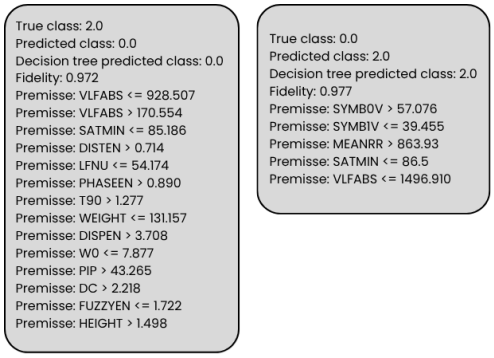

Fig. 2. Examples of Explanations for Misclassified Individuals in Moderate (2) and Severe (0) Classes

It is interesting to investigate the errors related to the Moderate and Severe classes, as most misclassifications occur due to confusion between these two categories. In Figure 2, we present explanations generated for a case in which the real class of the individual was Moderate but it was classified as Severe, as well as for a case in which the true class was Severe but it was predicted as Moderate. For comparison purposes, Figure 3 shows the beginning of some explanations for these two classes (Moderate and Severe) in which the algorithm correctly classified the instances.

In the first explanation of Figure 2, the algorithm begins by analyzing the VLFABS feature, as it does in most instances correctly classified as Moderate. The test of whether VLFABS is less than or equal to a threshold of approximately 930 proved to be common in these cases (see Figure 3). However, the algorithm subsequently evaluates a second threshold of VLFABS, which ends up leading the instance to a deeper leaf in the decision tree, corresponding to an incorrect classification. In the second explanation, the model generates a rather short set of premises, in which the saturation-related attributes—considered the most important—are analyzed only at the end which is the reason of the misclassification.

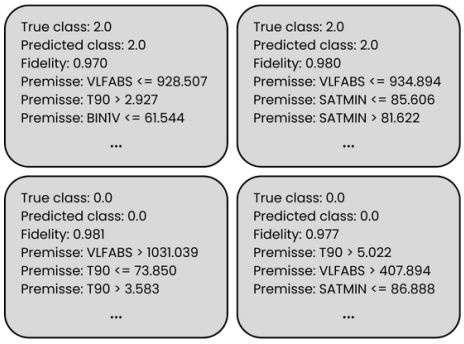

Fig. 3. Examples of explanations for correct classifications of individuals in the Moderate (2) and Severe (0) classes, the figure only shows the first three premises for the four individuals.

LORE also provides a set of counterfactual rules that can be analyzed to extract further insights from the model. For example, in Figure 4, we examine what changes would be necessary for the model to correctly classify as Moderate the first individual from Figure 2, which was incorrectly classified as Severe. The model did not provide any counterfactual rule for the second instance shown in Figure 2.

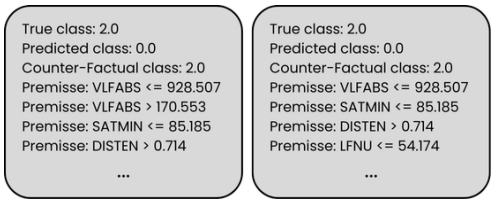

Fig. 4. Counterfactual generated for the individual explained on the first block of Figure 2, the counterfactual class indicates the classification generated by the counterfactual rule whose first four premises are displayed above

As we can see, the model provides two different sets of counterfactual rules that would result in the correct classification of the individual as belonging to the Moderate class. All of them start from the same threshold as in the original explanation; however, they subsequently evaluate different parameters that ultimately lead to the correct classification of the instance as Moderate.

As previously mentioned, the VLFABS parameter shows a significant correlation with the classes. This was already addressed in the work of [6], which used the Mean Decrease in Impurity to determine feature

importance in the RF. Here, we can make this importance more explicit by examining the frequency of this feature in both correct and incorrect classifications. To this end, we counted the most frequent features appearing in the explanations, and the results are presented in Figure 5. We also analyzed the most frequent features in the first, second, third, fourth, and fifth positions of the explanations, shown in Figure 6. As expected, the oxygen saturation indices are at the top of the list, confirming the attribute importance obtained through the Mean Decrease in Impurity, given that these features appear most frequently in the generated explanations.

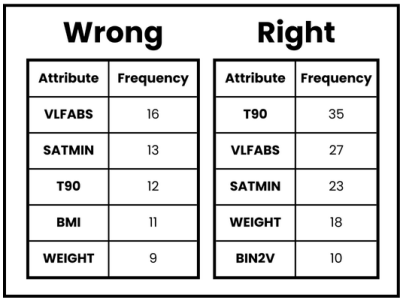

Fig. 5. Most frequent attributes in the cases where the RF algorithm was wrong and in the cases it was right.

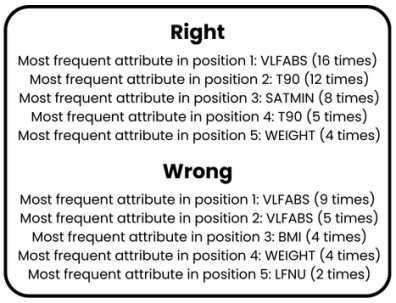

Fig. 6. Most frequent attributes at the first, second, third, fourth and fifth premises of the explanations.

Furthermore, VLFABS frequently appeared among the first positions in the explanations, confirming the trend previously identified, consistently ranking among the most relevant features in this analysis of patients with OSA. Although its physiological significance is not yet fully understood, the very low frequency (VLF) band has been associated with processes such as thermoregulation, activity of the renin-angiotensin-aldosterone system, and other humoral factors [9]. Studies have indicated that alterations in periodic breathing and oxygenation, common in apneic individuals, directly influence this component [10]. In addition, patients with OSA generally exhibit higher levels of VLF compared to healthy individuals. Additional evidence shows that VLFABS is an important feature for the automatic classification of OSA cases, reinforcing its value as a physiological marker [11].

As observed in the comparison between figures 1, 2, 3, and 4, LORE tends to generate similar or even identical rules across the explanations of different instances. We summarize in Table 2 the most frequently generated rules across correctly classified instances.

TABLE II
MOST FREQUENTLY GENERATED RULES FOR THE RF HITS.

Attribute	Operator	Threshold Range	Frequency
SATMIN	$\leq$	82.071–100.078	15
VLFABS	$\leq$	922.753–997.669	12
VLFABS	$>$	846.938–1058.446	8
BIN1V	$\leq$	61.070–73.669	7
WEIGHT	$\leq$	86.418–104.080	6
SATMIN	$>$	79.556–87.457	6
PERMEN	$\leq$	2.427–2.632	5
HEIGHT	$\leq$	1.748–1.828	5
AGE	$\leq$	31.228–31.438	4
GUZIKS	$>$	41.690–46.743	4

The data on the table above was generated using an algorithm that iterates through all the explanations produced by LORE and extracts the individual premises used in the construction of local rules. Each premise is decomposed into three components: feature, logical operator, and threshold value. All occurrences are compiled into a single dataset. The premises are then grouped based on the feature-operator pair, and their threshold values are sorted in ascending order. To account for similar rules that differ only by small numerical variations, we adopted a binning strategy: consecutive threshold values that differ by less than 10% are considered part of the same bin, and their frequencies are aggregated. This approach enables the identification of recurring decision regions rather than exact thresholds, which is more robust given the local and unstable nature of LORE explanations. The result is a table summarizing, for each feature and operator, the most frequently used threshold ranges along with their respective frequencies.

TABLE III
MOST FREQUENTLY GENERATED RULES FOR THE RF MISSES.

Attribute	Operator	Threshold Range	Frequency
VLFABS	$\leq$	922.753–971.670	9
SATMIN	$\leq$	78.409–90.992	7
GUZIKS	$>$	45.265–47.164	4
WEIGHT	$\leq$	92.163–102.630	4
SATMIN	$>$	85.186–91.332	4

In Table 2, we note the absence of the feature *T90*, which, despite being quite frequent and important feature (it was in fact the top-ranked feature among correctly classified instances), exhibits sparse thresholds such that

the frequency of any single rule involving it is low, as they are usually defined by varying unique values. For this attribute, we highlight the threshold range $\mathrm{T90} \leq 0.222$–0.250, which appears in four correctly classified instances. Table 3 presents the most frequent premises found in the misclassifications of the RF algorithm, organized in the same manner as Table 2, showing the most common threshold intervals. In this case we've chose to show only the top 5 rules because the frequencies of the others are bellow 4.

Among the threshold intervals shown above in Tables 2 and 3, we could highlight the premises $\mathrm{VLFABS} \leq 928.507$ and $\mathrm{VLFABS} > 1028.427$, which appeared five and three times, respectively, in exactly the same form across both correct and incorrect explanations.

B. Experiment 2: Complete Set of Examples

Here, explanations were generated for a model trained using all available data. Once again, the explanations are generated only for the real data and not for the synthetic data. The purpose here is to better understand how the final model, trained with a larger amount of data, classifies the examples. To this end, we provide the attributes most frequently analyzed in the explanations, which were this time counted separately for the four different classes, as shown in Table 4.

TABLE IV
MOST FREQUENT ATTRIBUTES FOR EACH CLASS IN THE MODEL TRAINED WITH ALL THE INSTANCES

SEVERE		**MILD**		**MODERATED**		**NON-APNEIC**	
Att.	**F**	**Att.**	**F**	**Att.**	**F**	**Att.**	**F**
SATMIN	106	VLFABS	61	VLFABS	61	T90	31
T90	101	SATMIN	57	T90	54	BMI	30
VLFABS	91	T90	56	SATMIN	52	VLFABS	28
ATTEN	56	WEIGHT	36	WEIGHT	28	AGE	26
SYMB2UV	42	AGE	23	ATTEN	20	SATMIN	20

Once again, the blood oxygen saturation indices are predominant, reaching the top 3 in almost all classes. Among the HRV indices, VLFABS remains highly relevant for data separation, especially in the Mild and Moderate classes. In the Severe and Non-Apneic classes, T90 seems to gain greater importance. However, as we will see next, the rules generated using T90 are again not very recurrent, in the sense that several different thresholds are created without concentrating in a narrower value range, unlike other features, e.g., VLFABS and SATMIN.

Furthermore, the most recurrent rules in each class were identified across all 291 explanations. These results can be seen in Tables 5-8, a threshold of 1% was used to determine that rules with the same attribute and operator would be regarded as similar and consequently grouped within the same threshold range, using the same methodology described before.

In the Severe class, we can frequently observe rules generated to check if SATMIN is less than a threshold ranging from 82.002 to 86.081, as well as rules to check if VLFABS is greater than a value between 1061.747 and 1072.482. This is quite interesting considering that studies indicate that the VLFABS of patients with OSA tend to have higher values in the very low frequency range [11].

TABLE V
MOST FREQUENTLY GENERATED RULES FOR THE FULL DATASET.

Attribute	Operator	Threshold Range	Frequency
SATMIN	$\leq$	77.738–86.081	117
SATMIN	$>$	80.580–86.620	60
VLFABS	$\leq$	928.507–971.670	45
PORTAS	$\leq$	53.258–55.846	23
PHASEEN	$\leq$	0.942–0.965	18
VLFABS	$>$	1061.747–1072.482	18
VLFABS	$\leq$	903.084–914.445	18
DISTEN	$\leq$	0.866–0.873	16
T90	$>$	0.137	15
PESO	$\leq$	98.300–102.500	14

TABLE VI
MOST FREQUENTLY GENERATED RULES FOR CLASS MILD (1).

Attribute	Operator	Threshold Range	Frequency
SATMIN	$>$	84.945–86.620	22
SATMIN	$>$	80.580–83.586	21
VLFABS	$\leq$	717.823–717.823	13
VLFABS	$\leq$	928.507–934.894	11
PORTAS	$\leq$	54.350–55.846	10
WEIGHT	$>$	82.163–82.429	7
PHASEEN	$\leq$	0.942–0.959	7
T90	$>$	0.154	7
VLFABS	$\leq$	905.556–914.445	7
T90	$>$	0.115	6

Again, the absence of T90 is justified by the fact that the thresholds generated for this attribute are more spread out, so they do not form significant clusters to be counted as recurrent. Among the relevant rules for the Severe class, we can highlight $\mathrm{T90} > 0.419$, with a frequency of 7 occurrences, and $\mathrm{T90} > 4.817$, with a frequency of 5 occurrences.

Finally, in the Non-Apneic patient class, recurrent checks of T90 are performed, always analyzing whether the value is below a very small threshold such as 0.075 or 0.063, as well as checks on age, weight, and BMI, patient characteristics that may be indicative of the presence of OSA. High BMI and weight greater than 80 kg are associated with classes 0, 1, and 2, showing correlation with severity. In this case, the check that minimum

saturation is greater than 90% was also quite recurrent, once again a behavior consistent with patients who do not have the disease.

TABLE VII
MOST FREQUENTLY GENERATED RULES FOR CLASS MODERATED (2).

Attribute	Operator	Threshold Range	Frequency
SATMIN	$\leq$	79.097–86.051	32
VLFABS	$\leq$	934.894–971.670	21
T90	$>$	0.137–0.137	10
VLFABS	$\leq$	903.084–914.445	8
PORTAS	$\leq$	53.601–54.555	6
SATMIN	$>$	80.771–80.936	4
BMI	$\leq$	31.247–31.872	4
PERMEN	$\leq$	2.541–2.565	4
WEIGHT	$\leq$	100.419–102.038	4
T90	$>$	1.717–1.728	3

TABLE VIII
MOST FREQUENTLY GENERATED RULES FOR CLASS NON-APNEIC (3).

Attribute	Operator	Threshold Range	Frequency
T90	$\leq$	0.075–0.075	8
SATMIN	$>$	90.998–91.685	8
BMI	$\leq$	30.967–31.465	5
BMI	$\leq$	26.405–26.451	5
AGE	$\leq$	25.353–25.463	4
T90	$\leq$	0.063–0.063	4
T90	$\leq$	0.115–0.115	4
SATMIN	$>$	86.152–86.231	3
SATMIN	$>$	89.940–89.940	3
WEIGHT	$\leq$	65.222–65.411	3

VI. CONCLUSIONS

We have successfully investigated the applicability of the LORE (Local Rule-based Explanations) method in the context of heart rate variability (HRV) for classifying the severity of sleep apnea. Using real-world data, we demonstrated that an explainability-based approach can not only maintain good predictive performance but also provide a clear interpretation of the model's decisions using a set of rules to explain different classifications.

The generation of local rules made it possible to identify relevant patterns associated with different severity classes, which can assist healthcare professionals in clinical decision-making. Couterfactuals were also very useful to understand what values and features should be modified for changing the predicted class, including to indicate what should be done to change the decision to the correct one when the black-box model misclassified the example. By identifying the most influential features and decision rules, our approach aids medical professionals to understand how HRV correlates with OSA severity and enhances trust in ML-driven diagnostic tools. The results reinforce the potential of transparent and interpretable artificial intelligence in sensitive medical applications.

As future work, we propose to include other XAI approaches for comparison, and further analyzing the clinical usefulness of the explanations generated, maybe in an evaluation of the VLFABS values. We should also investigate ways of improving LORE to generate more robust explanations. An initial attempt to generate more robust decision boundaries was made in LOREfs (LORE with fitness sharing) proposed by our research group.

REFERENCES

[1] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 80–89, 2018.

[2] "Lei geral de proteção de dados pessoais (lgpd) - lei nº 13.709, de 14 de agosto de 2018," https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm, 2018.

[3] A. B. Arrieta, N. Díaz-Rodríguez, J. D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-López, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[4] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Computing Surveys (CSUR)*, vol. 51, no. 5, pp. 1–42, 2018.

[5] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. San Francisco, CA, USA: ACM, 2016, pp. 1135–1144.

[6] R. R. dos Santos, M. B. Marumo, A. L. Eckeli, H. C. Salgado, L. E. V. Silva, R. Tinós, and R. F. Jr, "The use of heart rate variability, oxygen saturation, and anthropometric data with machine learning to predict the presence and severity of obstructive sleep apnea," *Frontiers in Cardiovascular Medicine*, vol. 12, p. 1389402, 2025.

[7] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "Local rule-based explanations of black box decision systems," *arXiv preprint arXiv:1805.10820*, 2018.

[8] XAI Library Contributors, "XAI Library: Explainable artificial intelligence tools," n.d., accessed: 2025-07-01. [Online]. Available: https://xai-tools.github.io/xai/

[9] F. Shaffer and J. P. Ginsberg, "An overview of heart rate variability metrics and norms," *Frontiers in Public Health*, vol. 5, p. 258, Sep 2017.

[10] D. P. Francis, L. C. Davies, K. Willson, P. Ponikowski, A. J. Coats, and M. Piepoli, "Very-low frequency oscillations in heart rate and blood pressure in periodic breathing: role of the cardiovascular limb of the hypoxic chemoreflex," *Clinical Science*, vol. 99, no. 2, pp. 125–132, Aug 2000.

[11] T. Shiomi, C. Guilleminault, R. Sasanabe, I. Hirota, M. Maekawa, and T. Kobayashi, "Augmented very low frequency component of heart rate variability during obstructive sleep apnea," *Sleep*, vol. 19, no. 5, pp. 370–377, Jun 1996.