IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 11 October 2023, accepted 8 January 2024, date of publication 18 January 2024, date of current version 25 January 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3355761

METHODS

Explainable Non-Contact Sleep Apnea Syndrome Detection Based on Comparison of Random Forests

IKO NAKARI, (Member, IEEE), AND KEIKI TAKADAMA, (Member, IEEE)
Department of Informatics, The University of Electro-Communications, Chofu, Tokyo 182-8585, Japan

Corresponding author: Iko Nakari (iko0528@cas.lab.uec.ac.jp)

This work was supported by the Grant-in-Aid for Japan Society for the Promotion of Science (JSPS) Research Fellow under Grant 22KJ1367.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ota Memorial Hospital Ethics Committee under Application No. 19000002.

ABSTRACT This paper focuses on Sleep Apnea Syndrome (SAS) and proposes the novel eXplainable AI (XAI) method that extracts characteristics of SAS by comparing the datasets of the SAS patients and the non-SAS subjects. For this issue, this paper (i) employs "two" Random Forests (RFs) to respectively learn the models for the SAS patients and the non-SAS subjects to classify the WAKE/non-WAKE stage, (ii) compares the two learned RFs to find their difference as the physiological characteristic of SAS, and (iii) proposes the SAS detection method based on the difference between the two learned RFs. Through the human subject experiment of the SAS detection based on the biological vibration data acquired from the mattress sensor during sleep, the following implications have been revealed: 1) RF learned from the SAS patient data classifies the WAKE/non-WAKE stage from the viewpoint of the "low" frequencies of the biological vibration data, while RF learned from the non-SAS subject data classifies it from the viewpoint of its "high" frequencies; and 2) the SAS patients have the WAKE stage with the low frequencies of the biological vibration data caused by disturbances in the autonomic nervous system due to apnea/hypopnea, while the non-SAS subjects do not have it but have the usual WAKE stage with the high frequencies caused by large body movements, which has a potential of the new characteristic of SAS instead of respiration as a traditional characteristic of SAS.

INDEX TERMS XAI, random forests, sleep apnea syndrome, feature importance, mattress sensor.

I. INTRODUCTION

The accumulation of sleep debt increases the risk of industrial and traffic accidents [1], [2], [3] and also increases the risk of developing diseases such as depression, dementia, and lifestyle-related diseases [4], [5], [6]. For these reasons, sufficient sleep is necessary for a productive daily life and a healthy life. Even in such sufficient sleep, however, sleep deprivation can be caused by sleep disorders. Sleep apnea syndrome (SAS) is one of the most common sleep disorders. SAS causes hypopnea (i.e., weakening of breathing) and apnea (i.e., cessation of breathing) during sleep, which worsens sleep quality. According to the global survey of obstructive sleep apnea syndrome (OSAS), the population of OSAS patients is estimated to be 78 million in the USA, 242 million in China, and 31 million in Japan [7]. Although the estimated number of patients is large, many of them are unaware of their suffering from OSAS. Furthermore, they unconsciously suffer from lifestyle-related diseases such as hypertension, myocardial infarction and so on [8], which increase the national cost of medical care [9]. From the facts, a regular diagnosis of SAS is necessary. The gold standard method of diagnosis needs a polysomnography (PSG) test which measures EEG, EOG, EMG and so on, and through the PSG test the sleep quality (sleep stage) is also defined based Rechtschaffen & Kales (R&K) method [10]. To define the severity of the SAS, the Apnea-Hypopnea Index (AHI) is employed, AHI counts apnea and hypopnea

The associate editor coordinating the review of this manuscript and approving it for publication was Nikhil Padhi.

events per sleep hour. An AHI score of 5-15 indicates mild, 15-30 moderate, and over 30 severe sleep apnea. This PSG test requires a subject to attach multiple sensors to his/her head and body and takes high cost which makes it difficult to get the PSG test regularly.

For this reason, the SAS detection methods have been developed. Concretely, most of them judge SAS according to the number of detected apnea/hypopnea from the respiration amplitude and the frequency analysis of the biological vibration data acquired from a mattress sensor [11], [12], [13]. However, it is difficult to detect apnea and hypopnea from a mattress sensor because labored breath (i.e., effort respiration) caused by apnea and/or hypopnea is very similar to normal respiration. Furthermore, it is also difficult to detect "hypopnea" from a mattress sensor in comparison with "apnea" because weak respirations in hypopnea are hard to be detected due to a similarity of normal respiration.

To tackle this problem, this paper focuses on the WAKE stage (i.e., shallow sleep) in the sleep stages as a new characteristic of SAS, instead of respiration as a traditional characteristic of SAS. This is because (1) the WAKE stage in the SAS patients often occurs in comparison with that in the non-SAS subjects due to apnea/hypopnea; (2) our previous research [14] found that it is difficult to detect the WAKE stage in the SAS patients by the Machine Learning (ML) model learned for the non-SAS subjects. These facts hypothesize that the characteristic of the WAKE stage in the SAS patients differs from the non-SAS subjects. To clarify such characteristics, this paper proposes the novel eXplainable AI (XAI) method that extracts characteristics of SAS by comparing a classification of WAKE/non-WAKE between the SAS patients and the non-SAS subjects. Concretely, this paper (i) employs "two" Random Forests (RFs) [15] as one of ML to respectively learn the models for the SAS patients and the non-SAS subjects to classify WAKE/non-WAKE using the biological vibration data acquired from the mattress sensors, (ii) compares the two learned RFs to find their difference as the characteristics of the WAKE stage in SAS, and (iii) proposes the SAS detection method based on the difference between the learned two RFs. We employ RF because of the following reasons: (1) the accuracy of RF is high thanks to the advantage of the ensemble method; (2) the rules in the decision tree constructed by RF are easier to be extracted than any other black-boxed models (e.g., deep learning); and (3) the feature importance, which measures how the features contribute to classifying training data, helps us to briefly understand what RF learned.

What should be noted here is that the proposed approach based on "two" RFs can provide more understanding of what RF learned than the conventional approach based on "one" RF which trains both data of the SAS patients and the non-SAS subjects at the same time. This is because the proposed approach based on "two" RFs makes it easy to identify the different tendencies of the SAS patients and the non-SAS subjects by comparing their feature importance, while the conventional approach based on "one" RF can calculate the feature importance but cannot tell us to understand how to classify SAS/non-SAS (see Section VI for detail). From this advantage, the proposed approach based on "two" RFs contributes to another understanding of SAS by extracting its characteristics from the biological vibration data acquired from mattress sensors.

This paper is organized as follows. The next section summarizes the related works. Section III describes the RF, and Section IV proposes the novel XAI method for extracting the characteristics of SAS. The experiment is conducted in Section V, and the results are discussed in Section VI. Finally, our conclusion is given in Section VII.

II. RELATED WORK

A. NON-CONTACT SAS DETECTION

1) MATTRESS SENSOR

Most of the methods with a mattress sensor focus on apnea (not hypopnea) and were designed to detect apnea from the respiration amplitude calculated from the biological vibration data (i.e., a pressure value acquired from a mattress sensor) or the spectrum analysis of the biological vibration data filtered for respiration [12]. However, it is difficult to apply these methods to mild SAS patients who often cause hypopnea because hypopnea is difficult to be detected due to weak respirations. As another approach, abnormal respiration is detected by applying principal component analysis (PCA) for the filtered respiration signal acquired from a mattress sensor [11]. However, labored breath (i.e., effort respiration) caused by apnea and hypopnea are hard to be detected as abnormal respiration due to a similarity of normal respiration.

2) MICROPHONE

A severity of OSAS is estimated by a Gaussian mixture regression based on the features extracted from the time and the spectra domains of the recorded audio acquired from the ambient microphones [16]. This work revealed that the correlation coefficient of the estimated AHI (Apnea Hypopnea Index) and the correct AHI was 89.2% with 7.35 events/hr error of AHI. However, this method cannot always detect SAS because the AHI error is too large to precisely detect SAS (e.g., mild SAS subjects (whose AHI $\geq$ 5) may be wrongly detected as a healthy subject (AHI<5) when AHI is estimated with 7.35 smaller than the correct AHI).

3) RADAR

The abnormal respiration is detected by the algorithm in SleepMinder [17] which estimates AHI by measuring the breathing and body movement of a subject in bed from data of the radio-frequency sensor. This work revealed that the estimated AHI and the correct AHI have a correlation of 91% and moderate subjects (AHI>15) can be detected with a sensitivity of 89% and a specificity of 92%. However, mild subjects (AHI>5) are hard to be detected due to a difficulty in estimating AHI with high accuracy (which a specificity is 46%).

B. XAI METHODS ON SAS DETECTION

Some recent studies applied the XAI method in SAS detection. One is the study that applied LIME (Local Interpretable Model-Agnostic Explanations) [18] to an apnea detection ML model which is trained with signals acquired from PSG test [19]. This approach aimed to identify which specific features and their ranges were most significant in predicting apnea, thereby enhancing the explainability of ML models in SAS diagnosis. Although the method provides what features and values make the ML models predict apnea for the input data, the method is not effective if the ML models cannot predict apnea correctly or if the features of the input data are difficult to interpret.

The other study employs Grad-CAM (Gradient-weighted Class Activation Mapping) [20] for the model composed with CNN and RNN and the model predicts apnea [21]. Grad-CAM is a visual method of showing the basis for predictions, and in this work, it visualizes what input data (i.e., the signal of airflow and SpO_2) are predicted as apnea. This method only visualizes the input data and doctors need to analyze the result in detail.

III. RANDOM FORESTS

RF [15] is an ensemble learning method composed of multiple decision trees as a weak classifier, and determines the output (i.e., classification result) by the majority vote of the classification results of decision trees.

A. ALGORITHM OF RF

FIGURE 1 shows the overview of RF, which is executed as follows: (1) the training datasets are generated by RF randomly sampling from the whole training dataset, (2) the decision trees are constructed according to their own training datasets, and (3) the output is determined by the majority classification results of the decision trees. In this research, Gini impurity [22] is employed to determine the condition in the nodes of the decision trees. The value of Gini impurity decreases when the ratio of the same label in the sampled data in the node increases. The learning process of RF is summarized as follows:

1) The training datasets S_j, where $j = 1 \ldots N_{tree}$, is generated from the whole training dataset S by the bootstrap sampling which allows to select the same data from S. Note that N_{tree} is the number of the decision trees to be constructed.

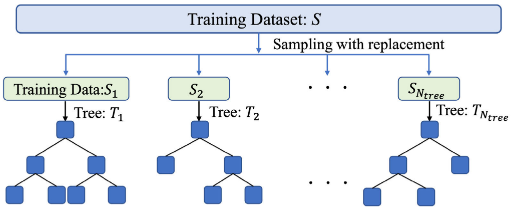


FIGURE 1. Overview of random forests.

2) The decision trees T_j, where $j = 1 \ldots N_{tree}$, are constructed according to S_j. In this construction, the following process is executed to generate the condition in the nodes of the decision tree.
   a) The different m_{try} features are randomly selected (by not allowing to select the same features). In the classification problem, it is recommended to set m_{try} as the square root of the total number of features.
   b) The feature with its threshold value is selected among many features with various threshold values to maximize the difference in Gini impurity before and after dividing the data. The condition of dividing data in the node is represented by the selected feature and its threshold value.
   c) For the divided data, repeat b) until reaching the pre-defined depth of the decision tree.

B. FEATURE IMPORTANCE

The feature importance is the contribution of the feature in classifying data. Concretely, the importance of the i-th feature y_i is calculated by an average of the difference in Gini impurity before and after dividing the data in all nodes. The formula of this calculation is shown in Eq. (1) as follows,

$$Imp(y_i) = \frac{\sum_{j=1}^{N_{tree}} \Delta(T_j(y_i))}{N_{tree}} \tag{1}$$

where $\Delta(T_j(y_i))$ indicates the difference of Gini impurity before and after dividing the data according to the feature y_i in the decision tree T_j. Since the feature importance of y_i increases when its impurity in the node decreases, the large feature importance means that the feature is important to classify the data.

IV. COMPARISON OF RFS FOR EXTRACTING SUBJECT CHARACTERISTICS

A. OVERVIEW

FIGURE 2 shows the overview of the proposed XAI method that extracts characteristics of SAS from the viewpoint of the WAKE stage as follows: (1) two RFs are independently trained from the datasets of the SAS patients and the non-SAS subjects, which are composed of the biological vibration data acquired from the mattress sensor and the correct label of WAKE/non-WAKE acquired from the PSG test (which determines the sleep stage from EEG, EOG and EMG acquired by attaching the electrodes to the body and head of the subjects); (2) two RFs are compared to find the difference rules in the generated decision trees between SAS/non-SAS. For example, the rules C and D show the difference between SAS/non-SAS, which rules respectively indicate the characteristics of SAS/non-SAS.

B. FEATURE BY POWER SPECTRUM

To extract characteristics of SAS by investigating the WAKE/non-WAKE stage, the power spectrum is employed

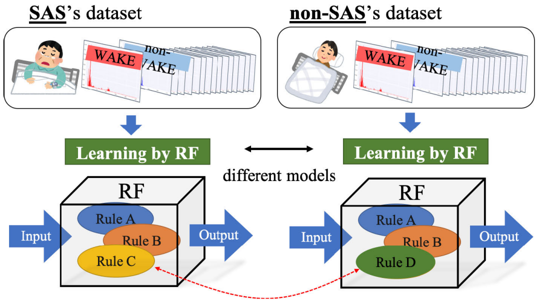


FIGURE 2. Overview of the comparison of RF.

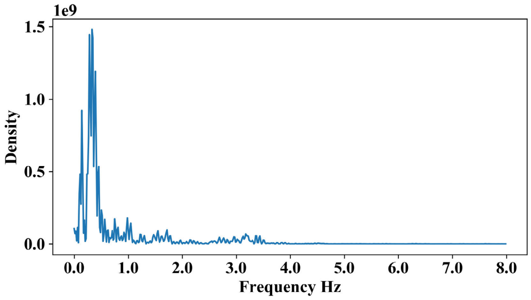


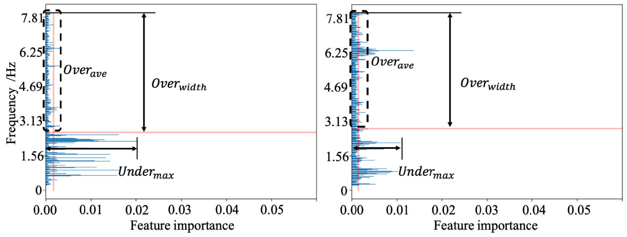
FIGURE 3. Power spectrum calculated from data of mattress sensor.

and calculated from the biological vibration data. This is because the raw data of the biological vibration data make it difficult to capture the characteristics of the vibration with heartbeat, respiration, and body movement, and the power spectrum can decompose these vibrations which is helpful to take account of all characteristics in the WAKE/non-WAKE stage. Concretely, the Fast Fourier Transform (FFT) [23] is applied to the biological vibration data with the 64-second window to convert it to the power spectrum. Note that, the apnea/hypopnea often lasts from 10 to 60 seconds in patients with mild and moderate SAS, a 60-second window is suitable and 64 (where the sampling frequency of the mattress sensor employed in this paper is 16 Hz and data size is $64 \times 16 = 1024$) was employed as the closest value in the factorial of 2 to which the FFT could be applied. Since the frequency of the data can be analyzed up to 8 Hz by FFT according to the sampling theorem [24], the data size of the power spectrum is 512 ($= 64 \times 8$) and the frequency resolution is 1/64Hz. FIGURE 3 shows the power spectrum calculated from the biological vibration data, where the vertical and horizontal axes indicate the density of the power spectrum and the frequency, respectively. In particular, the frequency band between 0.1 Hz and 0.3 Hz is related to respiration, and the frequency band between 0.6 Hz and 1.5 Hz is related to heart rate. Regarding the body movement, the density of the power spectrum becomes higher/lower as the body movement becomes larger/smaller. This power spectrum is calculated per second, and its frequency is employed for RF to learn a classification of the WAKE/non-WAKE stage.

C. EXTRACTION WHAT RF LEARNED BY FEATURE IMPORTANCE

To clarify the difference between SAS/non-SAS from the viewpoint of the WAKE stage, the proposed method employs the feature importance (described in section III-B.2) of the frequency of the biological vibration data of the SAS patients and the non-SAS subjects. FIGUREs 4(a) and 4(b) show the feature importance in the learned RFs of the SAS patient and the non-SAS subject, where the vertical and horizontal axes indicate the frequency and the feature importance, respectively. As shown in FIGURE 4, the distribution of the feature importance of the SAS patients is different from that of the non-SAS subject. Concretely, RF learned the low frequencies as the significant feature of the SAS patient while the high frequencies as the significant feature of the non-SAS subject. This implies that low frequencies (i.e., correspond to the vibrations of respiration and heart rate) are important for the SAS subjects when classifying the WAKE stage. Following this observation, Section VI-A presents a more detailed analysis highlighting the distinct feature importance patterns observed between SAS subjects and non-SAS subjects. Looking back on FIGURE 2, the rule C is regarded to classify the data according to the low frequencies while the rule D is regarded to classify the data according to the high frequencies.

To quantify such a difference, this paper proposed the index of Spectrum Importance Feature (SIF). To calculate SIF, the distribution of the feature importance is divided into the upper and lower sides according to the border of the frequency. This process corresponds to the two-class classification, and the border of the two classes is determined by the smallest Gini impurity when dividing the upper and lower sides. The detailed process is shown in Algorithm 1. The detailed process is shown in Algorithm 1. First, the feature importance (FI) of all frequencies determined by RF, is set to *Imp*[ ] (i.e., *Imp*[0] =FI of 1/64Hz, *Imp*[1]=FI of 2/64Hz, …, *Imp*[511] =FI of 512/64Hz(=8Hz)), and the classification threshold of the feature importance (which classifies feature importance into small or large) is set to *th*[ ] by dividing the average of the feature importance of all frequencies (aveFI) into 20 partitions (i.e., *th*[0] is aveFI*1/20, *th*[1] = aveFI*2/20, …, *th*[19] =aveFI*20/20) (lines 1 and 2). Note that *th*[ ] is indicated by the red vertical line in FIGURE 4 and the number of partitions (20) is determined by the pre-experiment. Next, the initial values of the maximum impurity difference (*maxDiff*) and the division frequency (*divideIndex*) are both set to 0 (lines 3 and 4). From the line 5, the optimal division frequency is searched from 1/10 * *Imp*.Length to 9/10 * *Imp*.Length to appropriately divide the upper and lower sides. Note that the division frequency is not searched for the first and last one-tenth of the total length because they are susceptible to noise. Given the current division frequency (i), the feature importances of all frequencies (*Imp*[k]) are classified into small or large by comparing with *th*[j] by varying *th*[j]. Concretely, if *Imp*[k] $<$ *th*[j] in the case of $k < i$ (i.e., the frequency k is smaller



(a) SAS (b) non-SAS

FIGURE 4. Feature importance of (a) SAS patient and (b) non-SAS subject.

than i, meaning the lower side), SL (i.e., the number of the feature importance which is Smaller than $th[j]$ in the Lower side) increases; otherwise LL (i.e., the number of the feature importance which is Larger than $th[j]$ in the Lower side) increases (lines 9 and 10). Similarly, if $Imp[k] < th[j]$ in the case of $k \geq i$ (i.e., the frequency k is same or larger than i, meaning the upper side), SU (i.e., the number of the feature importance which is Smaller than $th[j]$ in the Upper side) increases; otherwise LU (i.e., the number of the feature importance which is Larger than $th[j]$ in the Upper side) increases (lines 11 and 12). Note that SL, LL, SU, LU are set to 0 when the current division frequency (i) is updated (line 7). After counting SL, LL, SU, LU in all frequencies (k) in a certain i and j, the difference in Gini impurities before and after dividing into the lower and upper side ($currentDiff$) is calculated (line 15). This calculated difference of Gini impurity is compared every time to find the maximum difference (lines 16 to 18).

In FIGURE 4, the red horizontal line represents the borderline of dividing into the upper and lower sides of the distribution of the feature importance. After dividing the distribution horizontally, SIF is calculated by Eq.(2),

$$\mathrm{SIF} = \frac{Under_{max} \times Over_{width}}{Over_{ave}} \tag{2}$$

where $Under_{max}$ indicates the maximum feature importance of the lower side, $Over_{width}$ indicates the range of the upper side from the red borderline to the last frequency, and $Over_{ave}$ indicates the average of the feature importance in the upper side. Since the large feature importance in the SAS patients is shown together in the narrow range in the lower side (not shown separately in the wide range) in comparison with the non-SAS subjects, $Under_{max}$ and $Over_{width}$ increase in the SAS patients while $Over_{ave}$ increases in the non-SAS subjects. In detail, $Under_{max}$ becomes large in the SAS patients because the maximum size of the feature importance in the lower side in the SAS patients is larger than the non-SAS subjects. $Over_{width}$ becomes large in the SAS patients because the distribution of the large feature importance is shown in the narrow range of the lower side in the SAS patients while such a distribution is shown in the wide range in the non-SAS subjects. Finally, $Over_{ave}$ becomes large in the non-SAS subjects because of the same reason of $Over_{width}$. From this difference, the SIF value of the SAS patients becomes large while that of the non-SAS subjects becomes small.

Algorithm 1 Divide the Distribution of Feature Importance

```
 1: $Imp[\,]$: Feature importance of all frequencies
 2: $th[\,]$: Classification threshold of feature importance
 3: $maxDiff \Leftarrow 0$: Maximum impurity difference
 4: $divedeIndex \Leftarrow 0$: Division frequency
 5: for $i = Imp$.Length$/10$ to $9 * (Imp$.Length$)/10$ do
 6:     for $j = 0$ to $th$.Length do
 7:         $SL, LL, SU, LU \Leftarrow 0$
 8:         for $k = 0$ to $Imp$.Length do
 9:             if $k < i$ then
10:                 if $Imp[k] < th[j]$ then $SL$++else $LL$++
11:             else
12:                 if $Imp[k] < th[j]$ then $SU$++else $LU$++
13:             end if
14:         end for
15:         $currentDiff \Leftarrow$
            calculateImpurityDiff($SL, LL, SU, LU$)
16:         if $maxDiff < currentDiff$ then
17:             $maxDiff \Leftarrow currentDiff$
18:             $devideIndex \Leftarrow i$
19:         end if
20:     end for
21: end for
```

TABLE 1. Details of SAS subjects.

SAS ID	Severity	Num. of epoch	WAKE
A	moderate	912	51
B	moderate	977	146
C	moderate	865	174
D	moderate	953	66
E	mild	1031	140
F	mild	825	66
G	moderate	934	191
H	mild	954	48
I	mild	952	60

TABLE 2. Details of non-SAS subjects.

Non-SAS ID	Num. of epoch	WAKE
a	848	46
b	584	53
c	704	103
d	607	75
e	595	44
f	420	34
g	651	53
h	860	35
i	720	98

V. EXPERIMENT

To investigate the effectiveness of the proposed method, the human subject experiment was conducted with the nine SAS patients and the nine non-SAS subjects. In this experiment, the proposed method was compared with the single RF as the conventional method which is directly trained to classify SAS/non-SAS from the entire subjects. By comparing the proposed method and the single RF, this paper also investigates whether the proposed method can provide high SAS decision accuracy while ensuring interpretability. Note that the proposed method trained RFs for each subject to classify WAKE/non-WAKE, while the conventional method trained one RF for the entire subjects to classify SAS/non-SAS. In order to match the learning method of the proposed method, the conventional method also learns an epoch-by-epoch. For this issue, the power spectrum of the epoch (i.e., 30 seconds) with the correct label of WAKE/non-WAKE was employed as the training data in the proposed method, while that with the correct label of SAS/non-SAS was employed as the training data in the conventional method. This means that both classification results in the proposed method (i.e., WAKE/non-WAKE) and in the conventional method (i.e., SAS/non-SAS) are evaluated in a unit of one epoch. However, since the conventional method should be evaluated in a unit of one person (not one epoch), it classifies SAS when the ratio of the classification result of SAS in all epochs is greater than 0.5. For example, if one subject has 100 epochs and the RF in the conventional method classifies more than 50 epochs as SAS, then the subject is classified as SAS. Finally, both methods were evaluated by the leave-one-out cross-validation (i.e., the methods train the model with the data of the eight SAS patients or the eight non-SAS subjects and test the model with the remaining one SAS patient or one non-SAS subject), and continue this evaluation by changing the patent/subject to be excluded.

The parameters of RF in the proposed and conventional method are set as follows: (1) the maximum depth of the decision tree is 10; (2) the number of the decision tree is 300; (3) the number of features employed to construct the decision tree is 23 ($\fallingdotseq \sqrt{512}$) as described Section III-A. The following evaluation criteria are employed: accuracy, precision, recall and F-measure for the SAS detection.

A. DATASET

TABLE 1 shows the details of the SAS subjects, where the row of "SAS ID," "Severity," "Num. of epoch" and "WAKE" indicate the labeled ID of the SAS patients, the severe category of SAS, the number of epochs during one night, and the number of epochs labeled with the WAKE stage, respectively. These nine SAS subjects were originally suspected of having SAS and were actually diagnosed with mild to moderate SAS by a doctor through a PSG test. Note that this paper employs many mild SAS patients who are difficult to be detected as described in Section I and II TABLE 2 shows the details of non-SAS subjects, where the row of "Non-SAS ID" indicates the labeled ID of the Non-SAS subjects. These nine non-SAS subjects have no apnea/hypopnea symptoms during sleep and self-identified as healthy.

The biological data of the patients/subjects is measured by PSG and the mattress sensor is placed under the mattress of the bed. In this paper, TANITA sleep scan SL511 (Tokyo, Japan) with a sampling rate is 16 Hz was employed as the mattress sensor. After the sleep, the sleep stage is determined according to the R&K method based on EEG, EOG and EMG, and the biological vibration data is converted to the power spectrum. When conducting this experiment, the ethics community of Ota General Hospital approved this study in agreement with Helsinki's declaration. All the subjects were explained about our study by the hospital staff before the human subject experiment and signed their consent. For the protection of privacy, we only received the data from the hospitals, without the names and addresses of the subjects.

TABLE 3. Result of SAS detection by the conventional method.

SAS ID	SAS count	non-SAS count	SAS ratio
A	843	2	99.8%
B	792	11	98.6%
C	662	6	99.1%
D	856	6	99.3%
E	842	59	93.5%
F	714	7	99.0%
G	706	23	96.8%
H	858	3	99.7%
I	845	6	99.3%

TABLE 4. Result of non-SAS detection by the conventional method.

non-SAS ID	SAS count	non-SAS count	SAS ratio
a	19	752	2.5%
b	11	441	2.4%
c	12	613	1.9%
d	8	446	1.8%
e	6	513	1.2%
f	2	376	0.5%
g	14	585	2.3%
h	4	819	0.5%
i	4	636	0.6%

B. RESULTS

1) PROPOSED METHOD: COMPARISON OF RFS

FIGUREs 5 and 6 respectively show the feature importance of the frequency of the biological vibration data of the SAS patients and the non-SAS subjects in the proposed method, where the vertical and horizontal axes have the same meaning of FIGURE 4 and the alphabet in the upper right corner of each graph indicates the subject ID. As shown in FIGURE 5, the feature importance of the low frequencies tends to be large while that of the high frequencies tends to be small. As shown in FIGURE 6, on the other hand, the large and medium feature importance is shown separately in the whole frequencies.

FIGURE 7 shows the SIF value of all patients/subjects, where the vertical and horizontal axes indicate the SIF value and the IDs of patients/subjects, respectively. In detail, the red and blue bars respectively indicate the results of the SAS patients and non-SA subjects, and the black dotted line is a threshold for the SAS detection determined manually. In determining the threshold, this paper set it to be able to detect all SAS patients. As shown in FIGURE 7, all SAS patients can be completely separated from all non-SAS subjects when setting the appropriate threshold even with many mild patients in the dataset. This suggests that accuracy, precision, recall, and F-measure become 100%. Note that, this manually determined threshold will be changed due to the dataset, so that it is important to decide the optimal threshold by adding sufficient subjects.

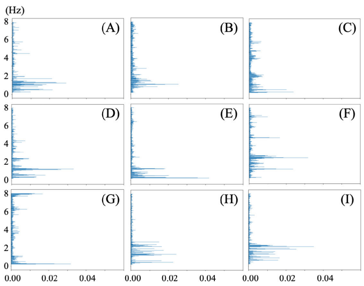


FIGURE 5. Feature importance of SAS patients trained with WAKE/non-WAKE.

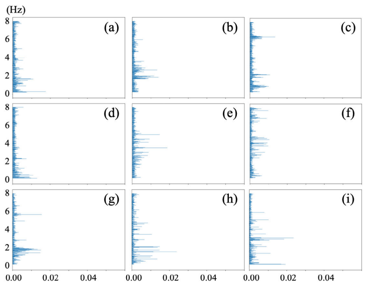


FIGURE 6. Feature importance of non-SAS subjects trained with WAKE/non-WAKE.

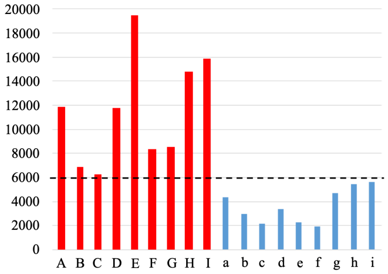


FIGURE 7. SIF values of SAS patients and non-SAS subjects.

2) NORMAL METHOD: SINGLE RF

TABLEs 3 and 4 respectively show the results of the SAS and non-SAS subjects in the conventional method, where the row of "SAS/non-SAS ID," "SAS count," "non-SAS count" and "SAS ratio" indicate the subject ID, the number of the epochs classified as SAS, the number of the epochs classified as non-SAS, and the ratio of the epochs classified as SAS and non-SAS, respectively. Considering that the conventional method classifies SAS when the SAS ratio is greater than 50.0%, TABLEs 3 and 4 show that all SAS patients are classified as SAS while all non-SAS subjects are classified as non-SAS. This suggests that accuracy, precision, recall, and F-measure are 100%.

VI. DISCUSSION

A. EXPLAINABILITY OF COMPARISONS OF RFS

To understand what RFs learned by the proposed method, the feature importance in the learned RFs is analyzed from the viewpoint of the spectrogram, which is represented by the three-dimensional values (i.e., time, frequency and power spectrum density (PSD)). FIGUREs 8 and 9 show the spectrograms of the non-SAS subject (ID: h) and SAS patient (ID: G) in one hour extracted from the whole sleep, where the vertical and horizontal axes indicate the frequency and the time, respectively. The color represents PSD which becomes dark/bright when its value becomes small/large. The orange solid line in the upper side of the figure indicates the sleep stage of PSG, and the orange/sky-blue color arrows indicate the WAKE stage with the large/small body movements which have the large/small PSD above 1Hz frequency, respectively. The green thin bar in the lower figure indicates the time when apnea/hypopnea occurs.

As shown in the non-SAS subjects shown in FIGURE 8, it is easy to classify WAKE/non-WAKE by the large/small PSD in the high frequencies (i.e., above 1 Hz frequency) occurred by the large/small body movement. In the SAS patients shown in FIGURE 9, on the other hand, it is difficult to classify WAKE/non-WAKE by the large/small PSD in the high frequencies because both the large and small body movements are found in the WAKE stage (i.e., the large body movements are found in the WAKE stage indicated by the orange color arrows, while the small ones are also found in the WAKE stage indicated by the sky-blue color arrows). Focusing on the small body movements in the WAKE stage, PSD around 0.3 Hz marked by the sky-blue circles tends to be large, and RF for the SAS patients learned this tendency while RF for the non-SAS subjects did not learn it. This suggests that the SAS patients have the different types of WAKE stages from the viewpoint of biological vibration data, while the non-SAS subjects do not have them.

This phenomenon can be explained from the viewpoint of the relationship between the autonomic nervous system and apnea/hypopnea. During sleep, the parasympathetic activity in the autonomic nervous system of the non-SAS subjects is generally stronger than the sympathetic activity. However, the sympathetic activity of the SAS patients becomes high during apnea/hypopnea, in order to promote breathing against hypoxia [25]. In such a situation, the sleep stage changes to the WAKE stage with the small body movement. FIGURE 9 shows this relationship that most of the WAKE stage with the small body movement (indicated by the sky-blue color arrows) appears when apnea/hypopnea (indicated by the green thin bars) occurs. Precisely, the WAKE stage with the small body movement starts to appear when the sympathetic activity becomes large, and then the apnea/hypopnea occurs. This is the reason why the small body movements in the WAKE stage are only found in the SAS patients. Furthermore, machine learning (i.e., RF in this paper) can learn the characteristics of WAKE in each subject by training each subject with a specialized model and the above facts have appeared in the feature importance distribution of each RF.

What should be noted here is that this implication cannot be found by the conventional method (i.e., the single RF) but only found by the proposed method which compares RFs for the SAS patients and the non-SAS subjects. The detailed implications are revealed as follows: (1) the WAKE stage with

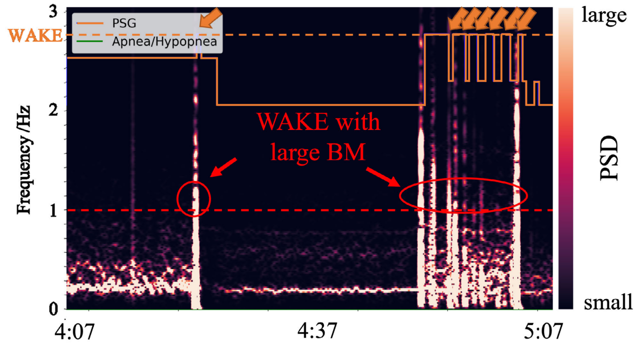


FIGURE 8. Spectrogram of the non-SAS subject (ID: h).

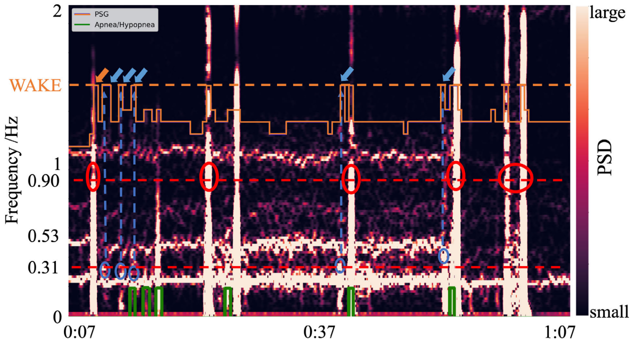


FIGURE 9. Spectrogram of the SAS subject (ID: G).

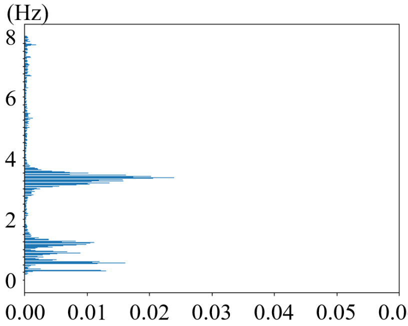


FIGURE 10. Feature importance extracted from RF trained with SAS/non-SAS for the entire subjects.

the small body movement (i.e, a high PSD in the low frequencies of the biological vibration data around 0.3 Hz) is only found in the SAS patients, which is caused by disturbance of the autonomic nervous system due to apnea/hypopnea; and (2) the WAKE stage with the large body movement is found in both the SAS patients and the non-SAS subjects. This suggests that the WAKE stage with the small body movement has a potential of the new characteristic of SAS instead of respiration as a traditional characteristic of SAS.

B. EXPLAINABILITY OF SINGLE RF

FIGURE 10 shows the feature importance of the frequency of the biological vibration data of the non-SAS subjects in the conventional method. This figure suggests that the frequency bands around 1 Hz and 3 Hz are effective in classifying SAS/non-SAS, but cannot tell us how RF classifies

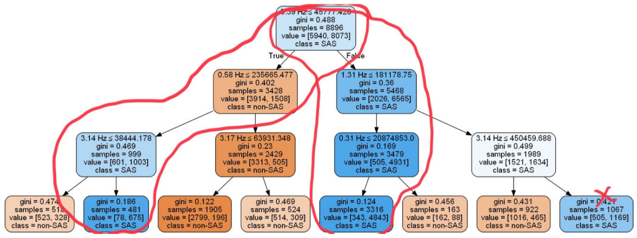


FIGURE 11. Example of decision trees in the learned RF.

SAS/non-SAS. To clarify this issue (i.e., to understand what RF learned), FIGURE 11 shows one of the decision trees in the learned RF, where the blue and orange squares show the rules for classifying SAS and non-SAS, respectively. The following rules are marked with the red lines in FIGURE 11.

- if PSD (0.39Hz) ≤ 45777 & PSD (0.58Hz) ≤ 235665 & PSD (3.14Hz) > 38444 then SAS
- if PSD (0.39Hz) > 45777 & PSD (1.31Hz) ≤ 181178 & PSD (0.31Hz) > 20874853 then SAS

The above rules can be roughly interpreted to classify SAS/non-SAS according to the low frequencies of the biological vibration data around 0.3 Hz (represented with the underline) in addition to around 1Hz and 3Hz, this suggests that it is important that the strength of vibration with around 0.3 Hz which corresponds to the respiration (note that when it is the normal respiration, the density of the power spectrum between 0.2 Hz to 0.26 Hz gets large). However, it is still difficult to grasp the tendency of the rules because of its complexity based on the multi-layered conditions. Furthermore, these rules are just two examples among all rules in all decision trees, which means that we cannot entirely understand what RF learned by analyzing a few rules. It is almost impossible to understand it as the number of trees increases and the depth of trees deepens. This is a fatal limitation of the single RF, although RF is categorized as an interpretable ML with a high expandability.

VII. CONCLUSION

This paper proposed the novel XAI method that extracted the characteristics of SAS by comparing RFs and investigated its effectiveness through the SAS detection based on the biological vibration data acquired from a mattress sensor. Concretely, the proposed method (i) learned the two RFs of the SAS patients and the non-SAS subjects independently to classify the WAKE/non-WAKE stage, (ii) compared the two learned RFs to find their differences as the physiological characteristics of SAS, and (iii) detected SAS according to the found difference of RFs.

Through the human subject experiment, the following implications have been revealed: (1) RF learned from the SAS patient data classifies the WAKE/non-WAKE stage from the viewpoint of the "low" frequencies of the biological vibration data, while RF learned from the non-SAS subject data classifies it from the viewpoint of its "high" frequencies; and (2) the SAS patients have the WAKE stage with the low frequencies of the biological vibration data caused by disturbances in the autonomic nervous system due to apnea/hypopnea, while the non-SAS subjects do not have it but have the usual WAKE stage with the high frequencies caused by large body movements, which suggests that the WAKE stage with the small body movement has a potential of the new characteristic of SAS instead of respiration as a traditional characteristic of SAS.

What should be noticed here is that the implications have only been obtained from the small number of the SAS patients and the non-SAS subjects, and therefore further careful qualifications and justifications, such as an increase of the patients/subjects, are needed to investigate the generality of our implications. Such important directions must be pursued in the near future in addition to (1) an investigation of whether the implications found in this paper are the same among mild, moderate and severe SAS patients; (2) the same analysis by estimating the WAKE/non-WAKE stage instead of providing the correct WAKE/non-WAKE stage in this paper.

REFERENCES

[1] P. Philip and T. Akerstedt, "Transport and industrial safety, how are they affected by sleepiness and sleep restriction?" *Sleep Med. Rev.*, vol. 10, no. 5, pp. 347–356, Oct. 2006.

[2] J. Kalsi, T. Tervo, A. Bachour, and M. Partinen, "Sleep versus non-sleep-related fatal road accidents," *Sleep Med.*, vol. 51, pp. 148–152, Nov. 2018.

[3] A. Kusztor, L. Raud, B. E. Juel, A. S. Nilsen, J. F. Storm, and R. J. Huster, "Sleep deprivation differentially affects subcomponents of cognitive control," *Sleep*, vol. 42, no. 4, Apr. 2019, Art. no. zsz016.

[4] N. Tsuno, A. Besset, and K. Ritchie, "Sleep and depression," *J. Clin. Psychiatry*, vol. 66, no. 10, pp. 1254–1269, 2005.

[5] J. M. Mullington, M. Haack, M. Toth, J. M. Serrador, and H. K. Meier-Ewert, "Cardiovascular, inflammatory, and metabolic consequences of sleep deprivation," *Prog. Cardiovascular Diseases*, vol. 51, no. 4, pp. 294–302, Jan. 2009.

[6] C. Holingue, A. Wennberg, S. Berger, V. Y. Polotsky, and A. P. Spira, "Disturbed sleep and diabetes: A potential Nexus of dementia risk," *Metabolism*, vol. 84, pp. 85–93, Jul. 2018.

[7] A. V. Benjafield, N. T. Ayas, P. R. Eastwood, R. Heinzer, M. S. M. Ip, M. J. Morrell, C. M. Nunez, S. R. Patel, T. Penzel, J.-L. Pépin, P. E. Peppard, S. Sinha, S. Tufik, K. Valentine, and A. Malhotra, "Estimation of the global prevalence and burden of obstructive sleep apnoea: A literature-based analysis," *Lancet Respiratory Med.*, vol. 7, no. 8, pp. 687–698, Aug. 2019.

[8] M. Okada, A. Takamizawa, K. Tsushima, K. Urushihata, K. Fujimoto, and K. Kubo, "Relationship between sleep-disordered breathing and lifestyle-related illnesses in subjects who have undergone health-screening," *Internal Med.*, vol. 45, no. 15, pp. 891–896, 2006.

[9] V. Kapur, D. K. Blough, R. E. Sandblom, R. Hert, J. B. de Maine, S. D. Sullivan, and B. M. Psaty, "The medical cost of undiagnosed sleep apnea," *Sleep*, vol. 22, no. 6, pp. 749–755, Sep. 1999.

[10] A. Rechtschaffen and A. Kales, *A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects*. Washington, DC, USA: U.S. Government Printing Office, 1968.

[11] S. H. Hwang, H. J. Lee, H. N. Yoon, D. W. Jung, Y. G. Lee, Y. J. Lee, D.-U. Jeong, and K. S. Park, "Unconstrained sleep apnea monitoring using polyvinylidene fluoride film-based sensor," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 7, pp. 2125–2134, Jul. 2014.

[12] M. L. Y. Davidovich, R. Karasik, A. Tal, and Z. Shinar, "Sleep apnea screening with a contact-free under-the-mattress sensor," in *Proc. Comput. Cardiol. Conf. (CinC)*, Sep. 2016, pp. 849–852.

[13] I. Castro, C. Varon, T. Torfs, S. Van Huffel, R. Puers, and C. Van Hoof, "Evaluation of a multichannel non-contact ECG system and signal quality algorithms for sleep apnea detection and monitoring," *Sensors*, vol. 18, no. 2, p. 577, Feb. 2018.

[14] I. Nakari, Y. Tajima, R. Takano, A. Toboru, and K. Takadama, "WAKE detection during sleep using random forest for sleep apnea syndrome patient," in *Proc. AAAI Spring Symp.*, 2019. [Online]. Available: https://ceur-ws.org/Vol-2448/

[15] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[16] E. Dafna, A. Tarasiuk, and Y. Zigel, "OSA severity assessment based on sleep breathing analysis using ambient microphone," in *Proc. 35th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2013, pp. 2044–2047.

[17] P. de Chazal, E. O'Hare, N. Fox, and C. Heneghan, "Assessment of sleep/wake patterns using a non-contact biomotion sensor," in *Proc. 30th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2008, pp. 514–517.

[18] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?' Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 1135–1144.

[19] A. R. Troncoso-García, M. Martínez-Ballesteros, F. Martínez-Álvarez, and A. Troncoso, "Explainable machine learning for sleep apnea prediction," *Proc. Comput. Sci.*, vol. 207, pp. 2930–2939, Jan. 2022.

[20] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[21] J. Jiménez-García, M. García, G. C. Gutiérrez-Tobal, L. Kheirandish-Gozal, F. Vaquerizo-Villar, D. Álvarez, F. del Campo, D. Gozal, and R. Hornero, "An explainable deep-learning architecture for pediatric sleep apnea identification from overnight airflow and oximetry signals," *Biomed. Signal Process. Control*, vol. 87, Jan. 2024, Art. no. 105490.

[22] R. Timofeev, "Classification and regression trees (CART) theory and applications," M.S. thesis, Center Appl. Statist. Econ., Humboldt Univ., Berlin, Germany, 2004, pp. 1–40.

[23] J. W. Cooley and J. W. Tukey, "An algorithm for the machine calculation of complex Fourier series," *Math. Comput.*, vol. 19, no. 90, pp. 297–301, 1965.

[24] C. E. Shannon, "Communication in the presence of noise," *Proc. IRE*, vol. 37, no. 1, pp. 10–21, Jan. 1949.

[25] K. Narkiewicz and V. K. Somers, "Sympathetic nerve activity in obstructive sleep apnoea," *Acta Physiologica Scandinavica*, vol. 177, no. 3, pp. 385–390, Mar. 2003.

IKO NAKARI (Member, IEEE) received the B.E. degree, in 2019, and the M.E. degree, in 2021. He is currently pursuing the Ph.D. degree in engineering with The University of Electro-Communications. His main research interests include health monitoring with wearable sensors, such as mattress sensors and smartwatches, particularly in sleep monitoring. His work focuses on a class imbalance of data, inter/intra-individual differences in machine learning. He is also motivated to discover new medical knowledge by machine learning from the viewpoint of computer science rather than medical science. He is a member of AAAI and Sleep- and AI-related research societies in Japan. He is a DC2 Research Fellow for young scientists with the Japan Society for the Promotion of Science, from 2022 to 2024.

KEIKI TAKADAMA (Member, IEEE) received the M.E. degree from Kyoto University, Japan, in 1995, and the Doctor of Engineering degree from The University of Tokyo, Japan, in 1998. From 1998 to 2002, he was a Visiting Researcher with the Advanced Telecommunications Research Institute (ATR) International. From 2002 to 2006, he was a Lecturer with the Tokyo Institute of Technology. In 2006, he moved to The University of Electro-Communications as an Associate Professor, where he has been a Professor, since 2011. His research interests include evolutionary computation, machine learning, healthcare systems, and sleep related issues. He is a member of ACM and major AI- and informatics-related academic societies in Japan. He served as the General Chair for the Genetic and Evolutionary Computation Conference (GECCO) 2018.