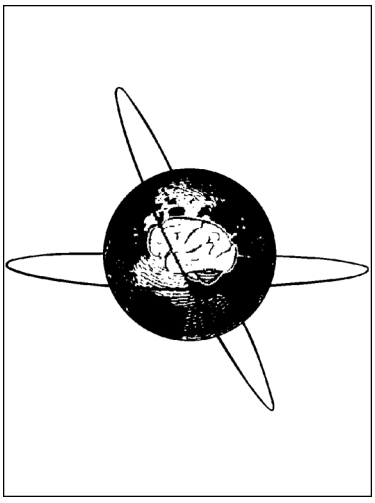
Journal Pre-proofs



R-R interval-based sleep apnea screening by a recurrent neural network in a large clinical polysomnography dataset

Ayako Iwasaki, Koichi Fujiwara, Chikao Nakayama, Yukiyoshi Sumi, Manabu Kano, Tetsuharu Nagamoto, Hiroshi Kadotani

PII: S1388-2457(22)00247-4
DOI: https://doi.org/10.1016/j.clinph.2022.04.012
Reference: CLINPH 2009935

To appear in: *Clinical Neurophysiology*

Received Date: 24 September 2021
Revised Date: 10 March 2022
Accepted Date: 12 April 2022

Please cite this article as: Iwasaki, A., Fujiwara, K., Nakayama, C., Sumi, Y., Kano, M., Nagamoto, T., Kadotani, H., R-R interval-based sleep apnea screening by a recurrent neural network in a large clinical polysomnography dataset, *Clinical Neurophysiology* (2022), doi: https://doi.org/10.1016/j.clinph.2022.04.012

R-R interval-based sleep apnea screening by a recurrent neural network in a large clinical polysomnography dataset

Ayako Iwasaki [1)]

Koichi Fujiwara (Corresponding Author), fujiwara.koichi@hps.material.nagoya-u.ac.jp [2) 3)]

Chikao Nakayama [2)]

Yukiyoshi Sumi [4)],

Manabu Kano [2)],

Tetsuharu Nagamoto [2)]

Hiroshi Kadotani [5)]

1) Faculty of Medicine, Kyoto University, Kyoto 606-8501, Japan

2) Department of Systems Science, Kyoto University, Kyoto 606-8501, Japan

3) Department of Materials Process Engineering, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, 464-8603, Japan

4) Department of Psychiatry, Shiga University of Medical Science, Shiga, 520-2192, Japan

5) Department of Sleep and Behavioral Sciences, Shiga University of Medical Science, 520-2192, Japan

Abstract

Objective: Easily detecting patients with undiagnosed sleep apnea syndrome (SAS) requires a home-use SAS screening system. In this study, we validate a previously developed SAS screening methodology using a large clinical polysomnography (PSG) dataset (N = 938).

Methods: We combined R-R interval (RRI) and long short-term memory (LSTM), a type of recurrent neural networks, and created a model to discriminate respiratory conditions using the training dataset (N = 468). Its performance was validated using the validation dataset (N = 470).

Results: Our method screened patients with severe SAS (apnea hypopnea index; AHI ≥ 30) with an area under the curve (AUC) of 0.92, a sensitivity of 0.80, and a specificity of 0.84. In addition, the model screened patients with moderate/severe SAS (AHI ≥ 15) with an AUC of 0.89, a sensitivity of 0.75, and a specificity of 0.87.

Conclusions: Our method achieved high screening performance when applied to a large clinical dataset.

Significance: Our method can help realize an easy-to-use SAS screening system because RRI data can be easily measured with a wearable heart rate sensor. It has been validated on a large dataset including subjects with various backgrounds and is expected to perform well in real-world clinical practice.

Keywords

Sleep apnea syndrome, Wearable sensor, Machine learning, Long short-term memory, Telemedicine.

Highlights

1. Sleep apnea syndrome (SAS) screening AI-based on R-R interval data was validated with a large clinical polysomnography dataset.
2. AUC of 0.92, a sensitivity of 0.80 and a specificity of 0.84 were achieved.
3. The SAS screening algorithm is easy to implement into a smartphone app.

Abbreviations

AHI: apnea hypopnea index

A/N: apnea/normal respiration

AS ratio: apnea sleep ratio

AUC: area under the curve

CPAP: continuous positive airway pressure

CSA: central sleep apnea

CVHR: cyclic validation of heart rate

ECG: electrocardiogram

EEG: electroencephalogram

EMG: electromyography

HRV: heart rate variability

LSTM: long short-term memory

OCST: out-of-center sleep testing

OSA: obstructive sleep apnea

PSG: polysomnography

RLS: restless leg syndrome

RNN: recurrent neural network

ROC: receiver operating characteristic

RRI: R-R interval

SAS: sleep apnea syndrome

SpO_2: saturation in peripheral oxygen

SUMS: Shiga University of Medical Science

TST: total sleep time

1. Introduction

Sleep apnea syndrome (SAS) is a disorder in which frequent apnea and hypopnea occur during sleep. Its severity is quantified in accordance with the apnea hypopnea index (AHI: respiratory events per hour of sleep), wherein SAS is an AHI of 5 or more, moderate SAS is 15 - 30, and severe SAS is 30 or more. Although SAS is a common disease, 80-90% of patients with SAS remain undiagnosed (Young et al, 1997). Continuous positive airway pressure (CPAP) has been shown to reduce the risk of lifestyle-related diseases (Jose et al, 2005), particularly in patients with AHI ≥ 30.

Many potential patients are undiagnosed because polysomnography (PSG), which is the gold standard test for SAS diagnosis, is costly and has limited access (Hassan and Haque, 2017). Instead, portable monitoring devices, which measure nasal pressure, chest and abdominal respiratory inductance plethysmography, and saturation in peripheral oxygen (SpO_2) (Kapur et al, 2017), have been used for SAS diagnosis (Kadotani et al, 2011). Although these methods and devices can be used at home, they require operational skills, and their diagnostic performance is not sufficiently high (Chesson et al, 2003). Thus, a simple and highly accurate SAS screening system that can be used easily at home is in need.

When apnea occurs during sleep, there is a decrease in SpO_2, which affects sympathetic nerve activities and induces changes in heart rate variability (HRV) (Somers et al, 1995; Qin et al, 2021;

Pathinarupothi et al, 2017). As demonstrated (Guilleminault et al, 1984), apnea periods are usually accompanied by short periods of respiration, in which higher heart rates are observed (Figure 1). During apnea periods, there are more evident fluctuations of heart rate compared to during normal respiration; this characteristic can be used to discriminate apnea from normal respiration (Figure 2). Apnea detection based on HRV has a great advantage over PSG and portable monitoring devices because HRV can be easily and accurately measured with a simple wearable device (Yamakawa et al, 2020). Based on this mechanism, we previously developed a method for detecting apnea using only electrocardiogram (ECG) data (Iwasaki et al, 2021). In our developed method, R-R intervals (RRIs) are extracted from ECG data and are used as input of long short-term memory (LSTM), which is a type of recurrent neural network. The trained model screened moderate-to-severe SAS patients with a sensitivity of 100% and a specificity of 100% in a clinical PSG dataset (N = 59), collected at Shiga University of Medical Science Hospital, Japan (Iwasaki et al, 2021). The Limitation of our previous study was that the number of subjects was not large enough to validate the algorithm and, accordingly, that the relationship between various complications with SAS and the performance could not be sufficiently investigated.

In this study, we evaluated the precise screening performance of our previously proposed model training procedure through its application to another large clinical PSG dataset (N = 938), collected

at the Nakamura clinic in Okinawa, Japan. Using a large number of subjects with various backgrounds, we further examined the impact of subject comorbidities on results.

2. Methods

We divided the clinical PSG dataset into a training dataset and validation dataset and retrained a machine learning model from the training dataset following our previously proposed method (Iwasaki et al, 2020). Finally, the screening performance of the retrained model was evaluated using the validation dataset.

2.1 Long short-term memory

Below, we briefly explain LSTM, which plays an important role in our SAS screening method.

Modern neural networks have been adopted in various fields, such as image analysis, text mining, and audio recognition, and a variety of network architectures have been developed. A recurrent neural network (RNN), focusing on time-series data analysis, receives output from a previous time point in addition to the current measurement. Thus, RNN can utilize past information as well as current information for time-series data analysis.

LSTM is a modification of RNN and can handle long-term dependencies by introducing a memory cell that holds long-term memory (Gers et al, 2000). As illustrated in Figure 3, the input, forget, and output gates can be trained to learn which information to store in the memory, for how

long, and when to read it out, respectively. Since LSTM can achieve a higher performance than the original RNN, LSTM has been used for speech recognition, natural language processing, and video analysis (Van Houdt et al, 2020). In our developed SAS screening method, LSTM is used for training an apnea/normal respiration (A/N) discriminant model, to which are input RRI data during sleep.

2.2 Sleep apnea screening method

The following is an overview of the screening method. The RRIs of subjects are extracted from an ECG signal in PSG data and divided into one-minute segments without overlap (see Figure 4). The length of the segment was determined based on the validation in our previous paper (Iwasaki et al, 2021). Since the raw RRI data are not sampled at equal intervals, the length of the vector is not constant. The PSG data were annotated with respiratory status every second, and segments containing apnea or hypopnea of more than x seconds was labeled as apneic (where x is an integer greater than or equal to zero), while others were labeled as normal respiration. x is set to 0 by default based on a previous study showing that heart rate changes not only during apnea periods but also before and after them (Vanninen et al, 1996). A model using LSTM is trained to determine whether the respiration condition of each RRI segment is apneic or normal (A/N) respiration. The trained A/N discriminant model is a network with three layers: an input layer, a hidden layer (LSTM with 32 units), and an output layer. It was trained using an Adam optimizer with a learning rate of 0.01,

batch size of 50, and run at 250 epochs. These hyperparameters of the LSTM model were determined by means of 5-fold cross-validation using the training dataset.

To determine whether a subject is a potential SAS patient, the apnea sleep ratio (AS ratio) A is calculated for each subject (Nakayama et al, 2015)

$$A = \frac{t_a}{t}$$

where t_a is the apneic periods determined by the model and t is the total sleep time (TST). A subject is considered to be a potential SAS patient if the AS ratio is greater than a predetermined threshold $\overline{A}$.

The AS ratio defined above requires the TST for each subject; however, measuring TST requires sleep scoring based on electroencephalogram (EEG) data analyzed by a technician. Thus, the average sleep latency in the training dataset was used instead to calculate the TST for simplification.

After training the LSTM-based A/N discriminant model, the RRI segments of the training dataset were input to the model to calculate the AS ratios. Using these AS ratios, we plotted a receiver operating characteristic (ROC) curve and defined the threshold of AS ratio $\overline{A}$ so that the Youden index (Youden 1950) is maximized. In addition, the AS ratio was calculated for each subject in the validation dataset. A subject was judged as a potential patient with SAS if the AS ratio was greater than $\overline{A}$.

2.3 Dataset

PSG recordings during sleep (6 - 7 hours) were collected from patients and healthy persons at the Nakamura clinic in Okinawa, Japan (N = 938). A PSG system (Alice6LDe or Alice5) included EEG, ECG (lead I or II, sampling frequency: 200 or 500 Hz), electromyography (EMG), SpO_2, chest and abdominal wall movements for respiratory efforts, nasal airflow, and a thermistor for respiratory monitoring. The study was approved by the Shiga University of Medical Science Research Ethics Committee (R2019-204).

After the removal of PSG data with strong artifacts in the ECG data, the dataset included 1015 subjects (original PSG dataset). Considering that the diagnostic criteria for obstructive sleep apnea (OSA) in children differ from those in adults (Seteia 2014; Berry et al, 2020), we excluded children under 12 years of age (N = 77) and created an adult PSG dataset (N = 938)

Each segment was labelled apnea or normal respiration based on the PSG data. The annotations were made by certified polysomnographic technologists of the Japanese Society of Sleep Research.

We extracted ECG data from PSG recordings and detected R waves using the Pan-Tompkins algorithm (Pan and Tompkins, 1985). RRIs were obtained from the detected R waves and divided into one-minute segments. Segments containing RRIs longer than 2,000 msec were deleted as invalid data. Then, each segment was normalized with a zero mean and a unit variance for each subject. In addition,

each RRI segment was labeled as apnea or normal respiration based on annotations in the PSG data by technicians.

To validate the screening performance of the model, we evaluated the results when it was applied to two other clinical datasets. One is a collection of PSG data from Shiga University of Medical Science Hospital (SUMS dataset, N = 59) whose subjects are Japanese, and the other is the Physionet apnea-ECG database (N = 69, Penzel et al, 2000).

2.4 Statistical Analysis

The training of the SAS screening model was conducted using Python 3.6.6 and TensorFlow 1.10.0. We used the Welch's t test for comparison of the ages between subjects who tested negative correctly and subjects with false positive. The significance level was set to $p < 0.05$, and computation was performed in Python 3.6.6 with SciPy 1.1.0. We calculated Spearman rank-order correlation coefficients between AHI and sleep parameters with Python 3.6.6 and Scipy 1.5.4.

3 Results

3.1 Screening performance

Subjects were classified into patients with severe SAS (AHI $\geq$ 30), moderate SAS (30 > AHI $\geq$ 15), or control (subjects with no or mild SAS: AHI < 15). Central and mixed apnea patients, as well as those with obstructive sleep apnea, were also included in this dataset. Subjects in the adult

PSG dataset (N = 938) were randomly split into training (N = 468) and validation (N = 470) datasets. A summarized profile and clinical characteristics of the subjects are shown in Tables 1 and 2, respectively. An example of ECG, RRI and respiratory status obtained from a patient is shown in Figure 1.

We built an LSTM model for screening severe SAS (AHI ≥ 30) from the training dataset and validated its performance through application to the validation dataset. The threshold of the AS ratio was 0.16. The ROC curves shown in Figure 5a, in which blue and orange lines indicate the ROC curve of the training and validation datasets, respectively. There is little difference between the two, which suggests that overfitting might not occur. Of the 470 subjects in the validation dataset, 163 subjects (35%) tested positive while 307 subjects (65%) tested negative. The model distinguished patients with severe SAS (N = 138) from subjects with AHI < 30 (N = 332) with an area under the curve (AUC) of 0.92, a sensitivity of 0.80, and a specificity of 0.84. 279 subjects (59%) with AHI < 30 correctly tested negative while 53 (11%) were false positives in the validation dataset. 110 severe SAS patients (23%) correctly tested positive while 28 (6%) were false negatives.

We also built an LSTM model for screening moderate-to-severe SAS (AHI ≥ 15) from the training dataset and validated its performance through application to the validation dataset. The threshold of the AS ratio was 0.13. The ROC curves are shown in Figure 5b; there is little difference

between the ROC curve of the training and validation datasets, which suggests that overfitting also might not occur in this model. In the validation dataset, the model distinguished patients with moderate-to-severe SAS (N = 221) from subjects with AHI < 15 (N = 249) with an AUC of 0.89, a sensitivity of 0.75, and a specificity of 0.87.

In addition, we built a model for mild SAS (AHI $\geq$ 5) in the same way; and the screening performance in the validation dataset was AUC of 0.83, sensitivity of 0.66, and specificity of 0.87.

3.2 Robustness of the performance

To validate the robustness of the trained severe SAS screening model (AHI $\geq$ 30), we randomly rearranged the training dataset and the validation dataset, trained models, and plotted the ROC for each trial (Figure 6). This procedure was repeated five times. The averages of the AUCs in the five trials were 0.92 $\pm$ 0.01, suggesting the steady performance of this method.

3.3 Change of the definition of apneic segments

In the above experiments, we labeled segments including apneas or hypopneas of one second or more as apneic. This may cause labeling of mostly normal breathing segments as apneic and normal breathing may be more likely to be identified as apneic. To explore this possibility, we modified the threshold to five seconds, which leads to a decrease in the number of apneic segments in the dataset by 4%. We retrained the model using the modified dataset and determined the subjects with AHI of

30 or higher, which resulted in an AUC of 0.93, sensitivity of 0.78, and specificity of 0.89 in the validation dataset. It is thought that, as a result of stricter judgments of apnea segments, the sensitivity of apnea detection is reduced.

3.4 Screening performance with children

In order to validate the performance when children were applied to this method, we constructed a model using the original PSG dataset. The profile of children added in the dataset is shown in Table 3. The method screened patients with severe SAS in the validation dataset with an AUC of 0.92, a sensitivity of 0.77, and a specificity of 0.87, and moderate or severe SAS with an AUC of 0.87, a specificity of 0.70, and a specificity of 0.88, which were slightly worse results than Figure 5.

3.5 Gender difference

In order to evaluate the effect of gender on the screening performance, we trained the model and validated the data separately for males (N = 715) and females (N = 223) with an AHI threshold of 30 (Figure 7). As a result, the validation data showed an AUC of 0.92, a sensitivity of 0.79, and a specificity of 0.84 for males (N = 364), and an AUC of 0.95, a sensitivity of 0.88, and a specificity of 0.84 for females (N = 106). The performance for females was improved compared to Figure 5a, while the performance for males was almost the same.

3.6 Validation using another clinical dataset

There may be a potential bias in the tendency of diagnosis and the attributes of the subject. To evaluate this point, we used another clinical dataset collected at Shiga University of Medical Science Hospital (SUMS), whose profile was shown in Table 4. When we applied the model trained from the Nakamura clinic data to the SUMS dataset, it was able to screen severe SAS (AHI $\geq$ 30) with an AUC of 0.94, a sensitivity of 0.93 and a specificity of 0.80, and moderate or severe SAS (AHI $\geq$ 15) with an AUC of 0.93, a sensitivity of 0.92 and a specificity of 0.89. These results suggest that our method is applicable to datasets collected at different institutions.

3.7 Validation using open datasets

Although the nationalities of the subjects were not recorded, most of the subjects in the dataset are thought to be Japanese. In order to investigate the effect of nationalities on the performance, we validated the developed method by using the Physionet apnea-ECG database (N = 69, Penzel et al, 2000). The subject profile of the database is summarized in Table 5. When we applied the threshold of AS ratio we determined using dataset collected at Nakamura clinic, it resulted in a sensitivity of 0.97 and a specificity of 0.58 with an AHI threshold of 30 and a sensitivity of 0.98 and a specificity of 0.74 with an AHI threshold of 15. When we determined the threshold of the AS ratio by using the half of the Physionet database (Table 5 (a), N = 34). As a result, the other half of the Physionet dataset

(Table 5 (b), N = 35) showed an AUC of 0.91, a sensitivity of 0.79 and a specificity of 0.95 with an AHI threshold of 30 and an AUC of 0.95, a sensitivity of 0.95 and a specificity of 0.86 with an AHI threshold of 15, which are comparable to results in Figure 5.

3.8 Relationship with sleep quality parameters

In order to investigate the relationship between AS ratio and actual sleep quality, the correlation coefficients between various sleep indices measured by PSG and the AS ratio were calculated in the validation dataset; the results are summarized in Table 6. The correlation coefficients between AS ratio and AHI and arousal index were 0.74 and 0.60, respectively. Figure 8 shows a scatter plot of AS ratio versus AHI and arousal index.

4. Discussion

The proposed method was able to screen severe SAS patients (AHI $\geq$ 30) with an AUC of 0.92, and moderate-to-severe SAS patients (AHI $\geq$ 15) with an AUC of 0.89. Table 7 summarizes the performance of the existing screening devices or methods validated using a large dataset of N > 100 (Mendonça et al, 2019; Álvarez et al, 2010; Huang et al, 2020, Roche et al, 2003, Gutiérrez-Tobal et al, 2015; Nakayama et al, 2019). Although SpO_2 measured with PSG data displays a relatively high performance to detect SAS (Álvarez et al, 2010), it has been reported that the performance of SpO_2-based methods declines when conducted in out-of-center sleep testing (OCST) (Ito et al, 2020). On the

other hand, RRI can be measured easily and precisely using an inexpensive wearable RRI sensor (Yamakawa et al, 2020), which realizes stable data collection in comparison with SpO_2, even in OCST. Table 3 shows that our method achieved results comparable with existing screening methods. Taking into consideration the fact that the number of subjects in this study is much larger than that of previous studies, our method is expected to stably exhibit high performance.

When screening severe SAS, 279 subjects with AHI < 30 correctly tested negative while 53 were false positives in the validation dataset. In the false positives, 23% of the subjects were over 60 years old, while 13% in the true negatives. This resulted in a statistical significance between the ages of the two groups ($p < 0.05$). It has been reported that HRV of an elderly person shows different patterns from that of a young or middle-aged person, since HRV decreases with age (Umetani et al, 1998) or is altered with mild cognitive impairment (Kong et al, 2020). These factors may prevent elderly persons from being screened correctly.

On the other hand, 110 severe SAS patients correctly tested positive while 28 were false negatives. Severe SAS patients with a history of arrhythmia account for 5% of the true positives (5 out of 110) and 11% of the false negatives (3 out of 28), suggesting the possibility that the existence of arrhythmia may lead to misclassification. The details in arrhythmia is in Supplementary Table 1. This result can be associated with the cyclic validation of heart rate (CVHR), in which the average or

variance of RRI fluctuates periodically during sleep apnea (Guilleminault et al, 1984). An arrhythmia may mask the cyclic validation pattern and make it difficult to appropriately detect apnea.

Diabetes also may affect the screening results. In the screening of severe SAS, there were five patients with diabetes in the subjects with false-negative results and three in the subjects with false-positive results. Diabetes can cause autonomic neuropathy, and it has been reported that the highly vulnerable parasympathetic nervous system is affected in the early stage of the disease, followed by sympathetic dysfunction (Pop-Busui 2010). This disturbance in the activity of the autonomic nervous system may alter CVHR patterns and cause misclassifications. It has been reported that CVHR did not occur in SAS patients with autonomic neuropathy including diabetes (Guilleminault et al, 1984).

There is the possibility that subjects with daytime sleepiness have sleep disorders other than SAS even if they suspect their SAS. Among these disorders, we examined restless leg syndrome (RLS) and narcolepsy. There were nine RLS patients with false-positive results in the validation dataset. It has been reported that HRV changes are observed in patients with not only RLS (Yıldız et al, 2018) but also periodic leg movements disorder which is a disorder closely related to RLS. Patients with these disorders show HRV with a pattern similar to that of CVHR (Hayano et al, 2011). The false positives may have been caused by these factors. In addition, there were three patients with narcolepsy in the validation dataset, all of whom currently tested negative. This result is consistent

with a previous study in which patients with narcolepsy were less likely to experience heart rate changes when the wake-up response occurred during sleep (Sorensen et al, 2013).

HRV is affected by gender as well as diseases and age; males are reported to have a lower heart rate than females and the distribution of their RRI is different (Voss et al, 2015). In order to evaluate such gender difference, we performed experiments separately (Figure 7). The performance for females was improved compared to Figure 5a, while the performance for males was almost the same. This may indicate that the male group has stronger heterogeneity than the female group. The number of males is larger than that of females, and the standard deviations of AHI in the male and female were 31.0 and 23.2, respectively, in this validation dataset. This diversity in males makes it harder for them to be screened correctly.

When we investigated the association of AS ratio with various sleep parameters, AHI and arousal index showed strong correlations therewith (Table 6). These results are consistent with the fact that apnea produces an arousal response (Eckert and Younes, 2014), which suggests that AS ratio is a good indicator of sleep quality. According to Table 6, there was no clear correlation with other sleep indices, such as wake after sleep onset, sleep efficacy, and total sleep time.

Our method exhibited high performance even when applied to another clinical dataset (SUMS dataset), suggesting that the trained model is applicable to data collected at other hospitals. We

further applied the model to a dataset whose subjects are not Japanese (Physionet apnea-ECG database) and got results with low specificity while the AUC was preserved. When we took into consideration that there are racial differences in heart rate variability (Hall et al, 2013) and determined the threshold of AS ratio using half of the Physionet database, the performance was comparable to the original results. These data indicate that the developed method exerts high performance even without re-training the model for each different race; we need to tune the threshold of the AS ratio for each race.

The limitation of this study is that the performance of the method deteriorates when the threshold of AHI is set to 5 for the screening of mild SAS. This may be because brief periods of apnea do not cause fluctuations in autonomic function, resulting in apnea and hypopnea segments with false negative. Another limitation is that the screening performance for patients with central sleep apnea (CSA) cannot be appropriately evaluated. Although two out of the three CSA patients in the validation dataset were screened correctly, it was difficult to evaluate them due to the very small number of patients with CSA. Since it has been reported that CVHR is clearly observed in CSA (Szollosi et al, 2007), the proposed method is expected to be capable of screening the patients without problem; however, further investigation is needed by collecting more data from CSA patients.

5. Conclusion

In this study, we validated our previously developed SAS screening method through application to a large clinical dataset. Our method was able to screen severe SAS (AHI $\geq$ 30) with an AUC of 0.92, and moderate or severe SAS (AHI $\geq$ 15) with an AUC of 0.89. In this study, we validated the proposed method using a large population with various backgrounds. In addition, we also confirmed the applicability of the model to populations of different institutions and nationalities.

While existing SAS screening devices based on SpO_2 were not appropriate for out-of-center testing due to the difficulty in signal measurement and analysis, our method is expected to solve the problems in the future because RRI data can be easily and stably measured using a wearable sensor (Yamakawa et al, 2013).

Thus, simple screening for SAS can be performed at home, which provides undiagnosed patients with an opportunity for diagnosis and proper treatment. We have developed a smartphone app implementing the SAS screening method, which can be connected to a wearable RRI sensor. In the future, we aim to perform a real-world prospective test using a wearable RRI sensor and the developed smartphone app.

Acknowledgements: All authors have seen and approved the manuscript. The work was performed at Department of Systems Science, Kyoto University, Kyoto 606-8501, Japan. This manuscript does not report on a clinical trial.

Financial support: This work was supported in part by JSPS KAKENHI #17H00872 and # 21H03851, and AMED #21uk1024005h0001.

Off-label or investigational use: None.

Conflict of interest: Koichi Fujiwara is with Quadlytics Inc as well as Nagoya University and Manabu Kano also is with Quadlytics Inc as well as Kyoto University. Other authors declare that the research was conducted in the absence of any commercial or financial relationship that could be construed as a potential conflict of interest.

Reference List

Álvarez D, Hornero R, Víctor Marcos J, Delcampo F. Multivariate analysis of blood oxygen saturation recordings in obstructive sleep apnea diagnosis. IEEE Trans Biomed Eng 2010;57:2816–24. https://doi.org/10.1109/TBME.2010.2056924.

Berry RB, Quan SF, Abreu AR, Bibbs ML, DelRosso L, Harding SM et al. The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications, version 2.6. Darien, Illinois. Am Acad Sleep Med 2020.

Chesson AL, Berry RB, Pack A. Practice parameters for the use of portable monitoring devices in the investigation of suspected obstructive sleep apnea in adults. Sleep 2003;26:907–13. https://doi.org/10.1093/sleep/26.7.907.

Eckert DJ, Younes MK. Arousal from sleep: implications for obstructive sleep apnea pathogenesis and treatment. J Appl Physiol 2014;116:302–13. https://doi.org/10.1152/japplphysiol.00649.2013.

Gers FA, Schmidhuber J, Cummins F. Learning to forget: Continual prediction with LSTM. Neural Comput 2000;12:2451–71. https://doi.org/10.1162/089976600300015015.

Guilleminault C, Winkle R, Connolly S, Melvinm K, Tilkian A. Cyclical variation of the heart rate in sleep apnoea syndrome. Mechanisms, and usefulness of 24 h electrocardiography as a screening technique. Lancet 1984;1:126–31.

Gutiérrez-Tobal G, Álvarez D, Gomez-Pilar J, del Campo F, Hornero R. Assessment of Time and Frequency Domain Entropies to Detect Sleep Apnoea in Heart Rate Variability Recordings from Men and Women. Entropy 2015;17:123–41. https://doi.org/10.3390/e17010123.

Hall MH, Middleton K, Thayer JF, Lewis TT, Kline CE, Matthews KA, et al. Racial differences in heart rate variability during sleep in women: The study of women across the nation sleep study. Psychosom Med 2013;75:783–90. https://doi.org/10.1097/PSY.0b013e3182a7ec5c.

Hassan AR, Haque MA. An expert system for automated identification of obstructive sleep apnea from single-lead ECG using random under sampling boosting. Neurocomputing 2017;235:122–30. https://doi.org/10.1016/j.neucom.2016.12.062.

Hayano J, Watanabe E, Saito Y, Sasaki F, Fujimoto K, Nomiyama T, et al. Screening for obstructive sleep apnea by cyclic variation of heart rate. Circ Arrhythmia Electrophysiol 2011;4:64–72. https://doi.org/10.1161/CIRCEP.110.958009.

Huang WC, Lee PL, Liu YT, Chiang AA, Lai F. Support vector machine prediction of obstructive sleep apnea in a large-scale Chinese clinical sample. Sleep 2020;43:1–11. https://doi.org/10.1093/sleep/zsz295.

Ito K, Uetsu M, Kadotani H. Validation of Oximetry for Diagnosing Obstructive Sleep Apnea in a Clinical Setting. Clocks & Sleep 2020;2:364–74. https://doi.org/10.3390/clockssleep2030027.

Iwasaki A, Nakayama C, Fujiwara K, Kadotani H, Sumi Y, Matsuo M, et al. Screening of sleep apnea based on heart rate variability and long short-term memory. Sleep Breath 2021;25:1821–1829. https://doi.org/10.1007/s11325-020-02249-0.

Jose M, Santiago J, Alvar GN, Marin JM, Carrizo SJ, Vicente E, et al. Long-term cardiovascular outcomes in men with obstructive sleep apnoea-hypopnoea with or without treatment with continuous positive airway pressure: an observational study. Lancet 2005;365:1046–53. https://doi.org/10.1016/S0140-6736(05)74229-X.

Kadotani H, Nakayama-Ashida Y, Nagai Y. Durability, safety, ease of use and reliability of a type-3 portable monitor and a sheet-style type-4 portable monitor. Sleep Biol Rhythms 2011;9:86–94. https://doi.org/10.1111/j.1479-8425.2011.00491.x.

Kapur VK, Auckley DH, Chowdhuri S, Kuhlmann DC, Mehra R, Ramar K, et al. Clinical practice guideline for diagnostic testing for adult obstructive sleep apnea: An American academy of sleep medicine clinical practice guideline. J Clin Sleep Med 2017;13:479–504. https://doi.org/10.5664/jcsm.6506.

Kong SDX, Hoyos CM, Phillips CL, McKinnon AC, Lin P, Duffy SL, et al. Altered Heart Rate Variability during Sleep in Mild Cognitive Impairment. Sleep 2020. https://doi.org/10.1093/sleep/zsaa232.

Mendonça F, Mostafa SS, Ravelo-García AG, Morgado-Dias F, Penzel T. A Review of Obstructive Sleep Apnea Detection Approaches. IEEE J Biomed Heal Informatics 2019. https://doi.org/10.1109/JBHI.2018.2823265.

Nakayama C, Fujiwara K, Matsuo M, Kano M, Kadotani H. Development of sleep apnea syndrome screening algorithm by using heart rate variability analysis and support vector machine. Annu Int Conf IEEE Eng Med Biol Soc. 2015;2015:8165-8. https://doi.org/10.1109/EMBC.2015.7320289.

Nakayama C, Fujiwara K, Sumi Y, Matsuo M, Kano M, Kadotani H. Obstructive sleep apnea screening by heart rate variability-based apnea/normal respiration discriminant model. Physiol Meas 2019:1–17. https://doi.org/10.1088/1361-6579/ab57be.

Pan J, Tompkins WJ. Real-Time Qrs Detection Algorithm. IEEE Trans Biomed Eng 1985;BME-32:230–6. https://doi.org/10.1109/tbme.1985.325532.

Pathinarupothi RK, Vinaykumar R, Rangan E, Gopalakrishnan E, Soman KP. Instantaneous heart rate as a robust feature for sleep apnea severity detection using deep learning. 2017 IEEE EMBS Int Conf Biomed Heal Informatics, BHI 2017 2017:293–6. https://doi.org/10.1109/BHI.2017.7897263.

Penzel T, Moody GB, Mark RG, Goldberges AL, Peter JH. The Apnea-ECG Database. Comput Cardiol 2000;27:255–8.Pop-Busui R. Cardiac autonomic neuropathy in diabetes: A clinical perspective. Diabetes Care 2010;33:434–41. https://doi.org/10.2337/dc09-1294.

Pop-Busui R. Cardiac autonomic neuropathy in diabetes: A clinical perspective. Diabetes Care 2010;33:434–41. https://doi.org/10.2337/dc09-1294.

Qin H, Keenan BT, Mazzotti DR, Vaquerizo-Villar F, Kraemer JF, Wessel N, et al. Heart Rate Variability during Wakefulness as a Marker of Obstructive Sleep Apnea Severity. Sleep 2021. https://doi.org/10.1093/sleep/zsab018.

Roche F, Pichot V, Sforza E, Court-Fortune I, Duverney D, Costes F, et al. Predicting sleep apnoea syndrome from heart period: A time-frequency wavelet analysis. Eur Respir J 2003;22:937–42. https://doi.org/10.1183/09031936.03.00104902.

Sateia MJ. International Classification of Sleep Disorders-Third Edition Highlights and Modifications. Chest 2014;146:1387–94. https://doi.org/10.1378/chest.14-0970.

Somers V K, Clary M P, Abboud F M, Somers Virend K, Dyken ME, Clary Mary P, et al. Sympathetic neural mechanisms in obstructive sleep apnea. J Clin Invest 1995;96:1897–904.

Sorensen GL, Knudsen S, Petersen ER, Kempfner J, Gammeltoft S, Sorensen HBD, et al. Attenuated heart rate response is associated with hypocretin deficiency in patients with narcolepsy. Sleep 2013;36:91–8. https://doi.org/10.5665/sleep.2308.

Szollosi I, Krum H, Kaye D, Naughton MT. Sleep Apnea in Heart Failure Increases Heart Rate Variability and Sympathetic Dominance. Sleep 2007;30:1509–14.

Umetani K, Singer DH, McCraty R, Atkinson M. Twenty-four hour time domain heart rate variability and heart rate: Relations to age and gender over nine decades. J Am Coll Cardiol 1998;31:593–601. https://doi.org/10.1016/S0735-1097(97)00554-8.

Van Houdt G, Mosquera C, Nápoles G. A review on the long short-term memory model. Artif Intell Rev 2020;53:5929–55. https://doi.org/10.1007/s10462-020-09838-1.

Vanninen E, Tuunainen A, Kansanen M, Uusitupa M, Länsimies E. Cardiac sympathovagal balance during sleep apnea episodes. Clin Physiol 1996;16:209–16. https://doi.org/10.1111/j.1475-097X.1996.tb00569.x.

Voss A, Schroeder R, Heitmann A, Peters A, Perz S. Short-Term Heart Rate Variability—Influence of Gender and Age in Healthy Subjects. PLoS One 2015;10:e0118308. https://doi.org/10.1371/journal.pone.0118308.

Yamakawa T, Fujiwara K, Kano M, Miyajima M, Suzuki Y, Maehara T, et al. Development of a wearable HRV telemetry system to be operated by non-experts in daily life. 2013 Asia-Pacific Signal Inf Process Assoc Annu Summit Conf APSIPA 2013 2013;1:1–4. https://doi.org/10.1109/APSIPA.2013.6694225.

Yamakawa T, Miyajima M, Fujiwara K, Kano M, Suzuki Y, Watanabe Y, et al. Wearable Epileptic Seizure Prediction System with Machine-Learning-Based Anomaly Detection of Heart Rate Variability. Sensors 2020;20:3987. https://doi.org/10.3390/s20143987.

Yıldız A, Yıldız C, Karakurt A. Assessment of cardiac autonomic functions by heart rate variability in patients with restless leg syndrome. Turk Kardiyol Dern Ars 2018;46:191–6. https://doi.org/10.5543/tkda.2018.33896.

Youden WJ. Index for rating diagnostic tests. Cancer 1950;3:32–5. https://doi.org/10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3.

Young T, Evans L, Finn L, Palta M. Estimation of the clinically diagnosed proportion of sleep apnea syndrome in middle-aged men and women. Sleep 1997;20:705–6. https://doi.org/10.1093/sleep/20.9.705.

Figures

Figure 1. An example of an electrocardiogram and respiratory waves from a patient (30 years old, apneal-hypopnea index; AHI = 86).

Figure 2. Whole R-R interval (RRI) data collected from (a) healthy subject (20 years old, apnea-hypopnea index; AHI = 0.9) and (b) patient (male, 46 years old, AHI = 22.2). (c) and (d) are their zoomed RRI data. The colored bands denote periods including a lot of apnea or hypopnea events.

Figure 3. An internal state of long short-term memory (LSTM) at timepoint t. LSTM receives the output from the previous point $h_{t-1}$to handle time-series as well as the current measurement x_t. In addition, LSTM introduces a cell memory C to handle long-term dependencies. The input gate i, the forget gate f, and the output gate o can be trained to learn which information to store in the memory, how long, and when to read it out, respectively. g is a new memory added to the memory cell.

Figure 4. Feature extraction framework. The R-R interval (RRI) is extracted from electrocardiogram (ECG) data and then split into periods of one-minute. The numbers in the figure indicate RRIs expressed in millisecond. Each interval is labeled as apnea (1) or normal respiration (0) based on the

annotations made by technicians, and an input vector of the screening model is built by clipping the RRI data.

Figure 5. Diagnostic performance of the Apnea/Sleep ratio (AS ratio) to detect severe (apnea-hypopnea index; AHI ≥ 30) (a) or moderate-to-severe (AHI ≥ 15) sleep apnea (b). Receiver operating characteristics curves (ROC) were constructed for training (N = 468, blue line) and validation (N = 470, orange line) datasets.

Figure 6. Robustness of the proposed model to detect severe sleep apnea (apnea-hypopnea index; AHI ≥ 30). The training and validation datasets were randomly replaced five times and a receiver operating characteristics curve (ROC) was plotted for each trial using the validation dataset (N = 470).

Figure 7. Comparison between the classification performance when males (a) and females (b) were trained and validated separately. Receiver operating characteristics curves (ROC) were constructed for training (blue line) and validation (orange line) datasets.

Figure 8. Scatter plot of Apnea/Sleep ratio (AS ratio) versus apnea-hypopnea index (AHI) (a) and arousal index (b). The red line indicates the regression line where AHI or arousal index is a predictor variable and AS ratio is a response variable. “r” represents a correlation coordinate between the two variables.

Tables

Table 1: Subject profile in the adult polysomnography dataset

(a) Training dataset (N = 468)

	Male			Female		
Age	AHI 0-15	15-30	30-	0-15	15-30	30-
13-30	48	6	7	18	0	1
31-60	98	53	81	45	8	7
61-	19	17	22	20	5	13

(b) Validation dataset (N = 470)

	Male			Female		
Age	AHI 0-15	15-30	30-	0-15	15-30	30-
13-30	38	4	8	15	1	0
31-60	125	50	94	44	6	9
61-	13	12	20	14	10	7

Age and sleep apnea severity (apnea-hypopnea index; AHI) distribution in the training and validation datasets collected at Nakamura clinic in Okinawa, Japan.

Table 2: Clinical and demographic characteristics of the subjects in the adult polysomnography dataset

	Training (N = 468)	Validation (N = 470)	All subjects (N = 938)
Apnea hypopnea index	25.3 (28.5)	25.5 (29.8)	25.4 (29.2)
Age -- yr	46.4 (16.2)	46.7 (14.9)	46.6 (15.6)
Female sex -- no. (%)	117 (25.0)	106 (22.6)	223 (23.8)
Body mass index (kg/m^2)	27.7 (5.3)	28.3 (5.1)	28.0 (5.2)
Hypertension -- no. (%)	125 (26.7)	128 (27.2)	253 (27.0)
Myocardial infarction -- no. (%)	3 (0.6)	1 (0.2)	4 (0.4)
Arrhythmia -- no. (%)	21 (4.5)	20 (4.3)	41 (4.4)
Diabetes -- no. (%)	39 (8.3)	41 (8.7)	80 (8.5)
Depression -- no. (%)	14 (3.0)	16 (3.4)	30 (3.2)
Allergic rhinitis -- no. (%)	45 (9.6)	50 (10.6)	95 (10.1)
Asthma -- no. (%)	43 (9.2)	34 (7.2)	77 (8.2)
COPD -- no. (%)	9 (1.9)	11 (2.3)	20 (2.1)
Restless leg syndrome -- no. (%)	52 (11.1)	50 (10.6)	102 (10.9)
Migraine -- no. (%)	4 (0.9)	8 (1.7)	12 (1.3)
OAB -- no. (%)	0 (0.0)	2 (0.4)	2 (0.2)

COPD: chronic obstructive pulmonary disease, OAB: overactive bladder.

Table 3: Subject profile of children used in the additional analysis (N = 77)

	Male			Female		
Age	AHI 0-15	15-30	30-	0-15	15-30	30-
0-5	24	4	2	9	2	1
6-12	20	1	2	11	0	1

Age and sleep apnea severity (apnea-hypopnea index; AHI) distribution (under 13) collected at Nakamura clinic in Okinawa, Japan.

Table 4: Subject profile of the SUMS dataset (N = 59)

	Male			Female		
Age	AHI 0-15	15-30	30-	0-15	15-30	30-
13-30	7	0	1	15	0	1
31-60	7	5	7	6	0	0
61-	0	4	4	0	1	1

Age and sleep apnea severity (apnea-hypopnea index; AHI) distribution (under 13) collected at Shiga University of Medical Science Hospital (SUMS) in Otsu, Japan.

Table 5: Subject profile of the Physionet database (N = 69)

(a) Dataset used to determine the threshold of AS ratio (N = 34)

	Male			Female		
Age	AHI 0-15	15-30	30-	0-15	15-30	30-
13-30	1	0	0	2	0	0
31-60	8	3	17	2	0	0
61-	0	1	0	0	0	0

(b) Validation dataset (N = 35)

	Male			Female		
Age	AHI 0-15	15-30	30-	0-15	15-30	30-
13-30	1	0	0	4	2	0
31-60	7	6	12	2	0	0
61-	0	1	0	0	0	0

Age and sleep apnea severity (apnea-hypopnea index; AHI) distribution of the Physionet apnea-ECG database (Penzel et al, 2000).

Table 6: Correlation coordinates between Apnea/sleep ratio (AS ratio) and sleep parameters

Sleep parameters	Correlation coordinates	P value
AHI	0.74	< 0.001
Arousal index	0.60	< 0.001
WASO	0.23	< 0.001
Sleep efficiency	-0.23	< 0.001
Total sleep time	-0.21	< 0.001

AHI: apnea hypopnea index, WASO: wake time after sleep onset.

Table 7: Summary of the performances of various screening methods using a large dataset

Ref	Number of subjects	Ages	Male sex (%)	Discrimination algorithm	Modality	Signal	AUC (Threshold of AHI)
Álvarez et al (2010)	148	52.9 ± 14.1	78	Logistic regression	PSG	SpO_2	0.97 (10)
Huang et al (2020)	6875	47.8 ± 14.5	76	Support vector machine	Questionnaire	None	0.82 (5), 0.80 (15), 0.78 (30)
Roche et al (2003)	147	53.8 ± 11.2	69	Classification and Regression Trees	PSG	ECG	0.76 (10)
Gutiérrez-Tobal et al (2015)	188	-	71	Logistic regression	PSG	ECG	0.89 (10)
Nakayama et al (2019)	115	42.2 ± 15.4	73	Classification and Regression Trees	PSG	ECG	0.84 (15)
Proposed	**938**	**43.5 ± 18.5**	**76**	**LSTM**	**PSG**	**ECG**	**0.89 (15), 0.92 (30)**

AUC: area under the curve, AHI: apnea hypopnea index, PSG: polysomnography, SpO_2: saturation of peripheral oxygen, ECG: electrocardiogram, LSTM: long short-term memory.

Objectives: Easily detecting patients with undiagnosed sleep apnea syndrome (SAS) requires a home-use SAS screening system. In this study, we validate a previously developed SAS screening methodology using a large clinical polysomnography (PSG) dataset (N = 938).

Methods: We combined R-R interval (RRI) and long short-term memory (LSTM), a type of recurrent neural networks, and created a model to discriminate respiratory conditions using the training dataset

(N = 468). Its performance was validated using the validation dataset (N = 470).

Results: Our method screened patients with severe SAS (apnea hypopnea index; AHI ≥ 30) with an area under the curve (AUC) of 0.92, a sensitivity of 0.80, and a specificity of 0.84. In addition, the model screened patients with moderate/severe SAS (AHI ≥ 15) with an AUC of 0.89, a sensitivity of 0.75, and a specificity of 0.87.

Conclusions: Our method achieved high screening performance when applied to a large clinical dataset.

Significance: Our method can help realize an easy-to-use SAS screening system because RRI data can be easily measured with a wearable heart rate sensor. It has been validated on a large dataset including subjects with various backgrounds and is expected to perform well in real-world clinical practice.

Figure 1

Click here to access/download;Figure;Figure 1.pdf

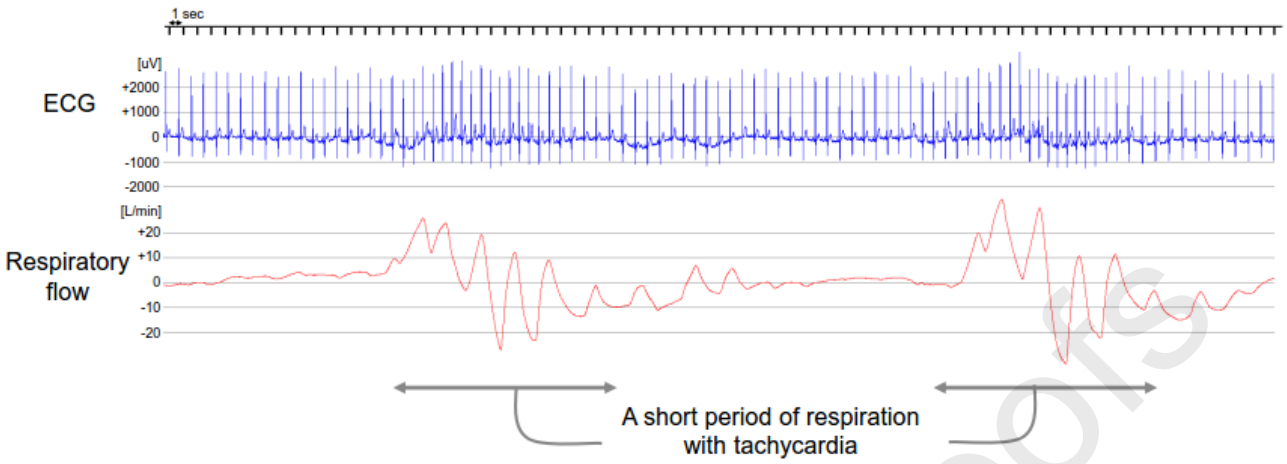


Figure 2

Click here to access/download;Figure;Figure 2.pdf

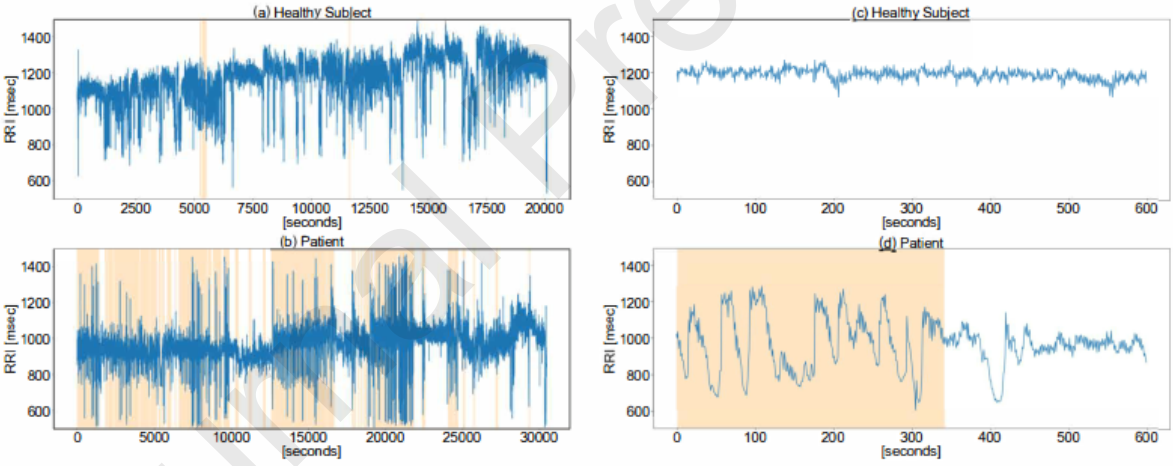


Figure 3

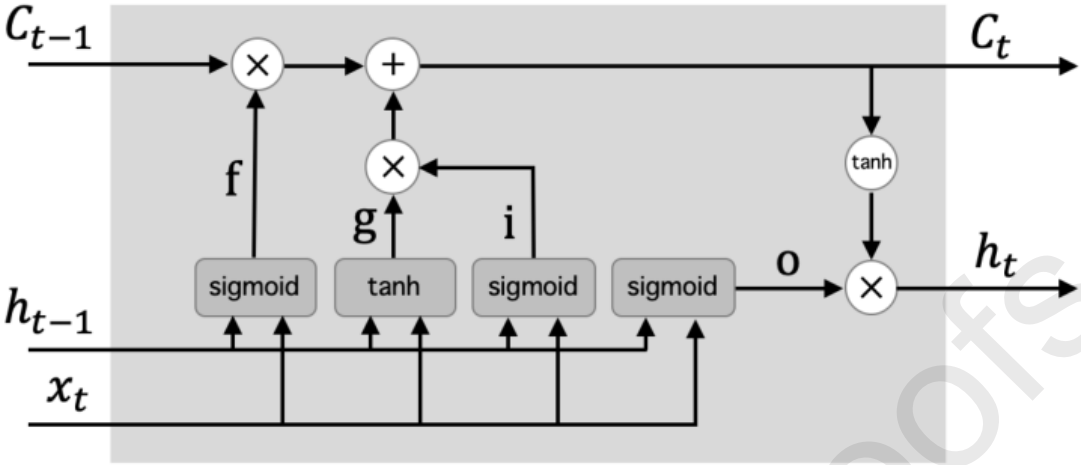


Figure 4

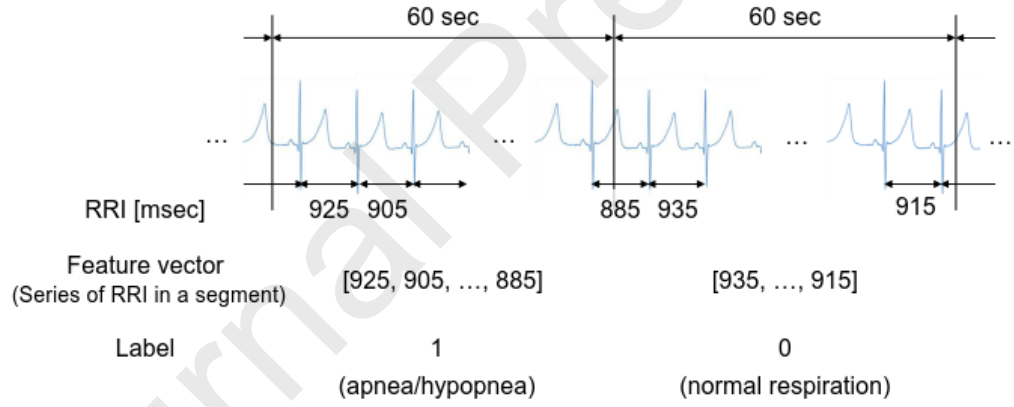


Figure 5

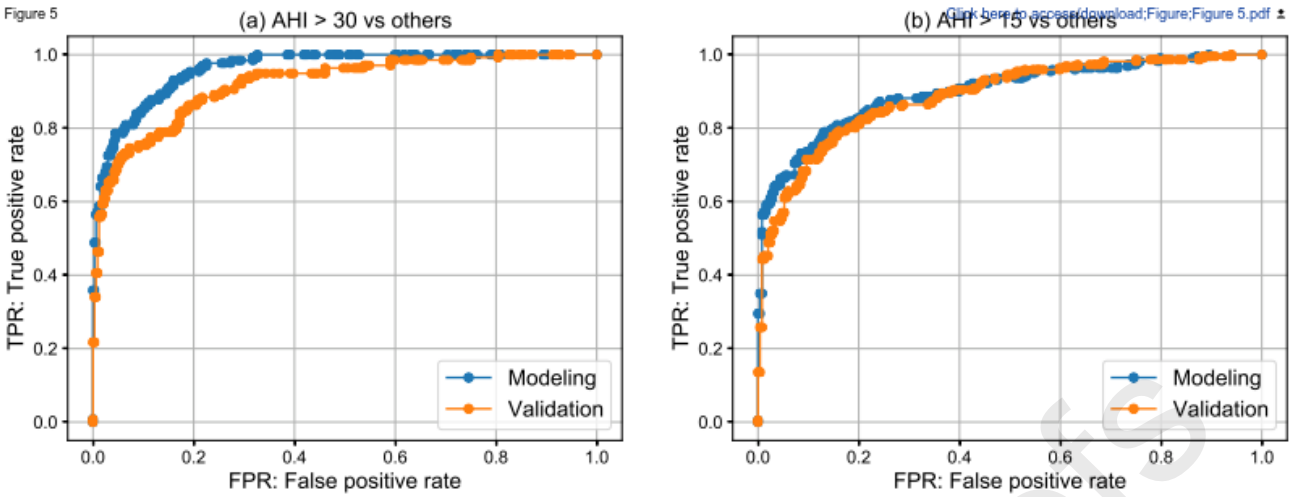


Figure 6

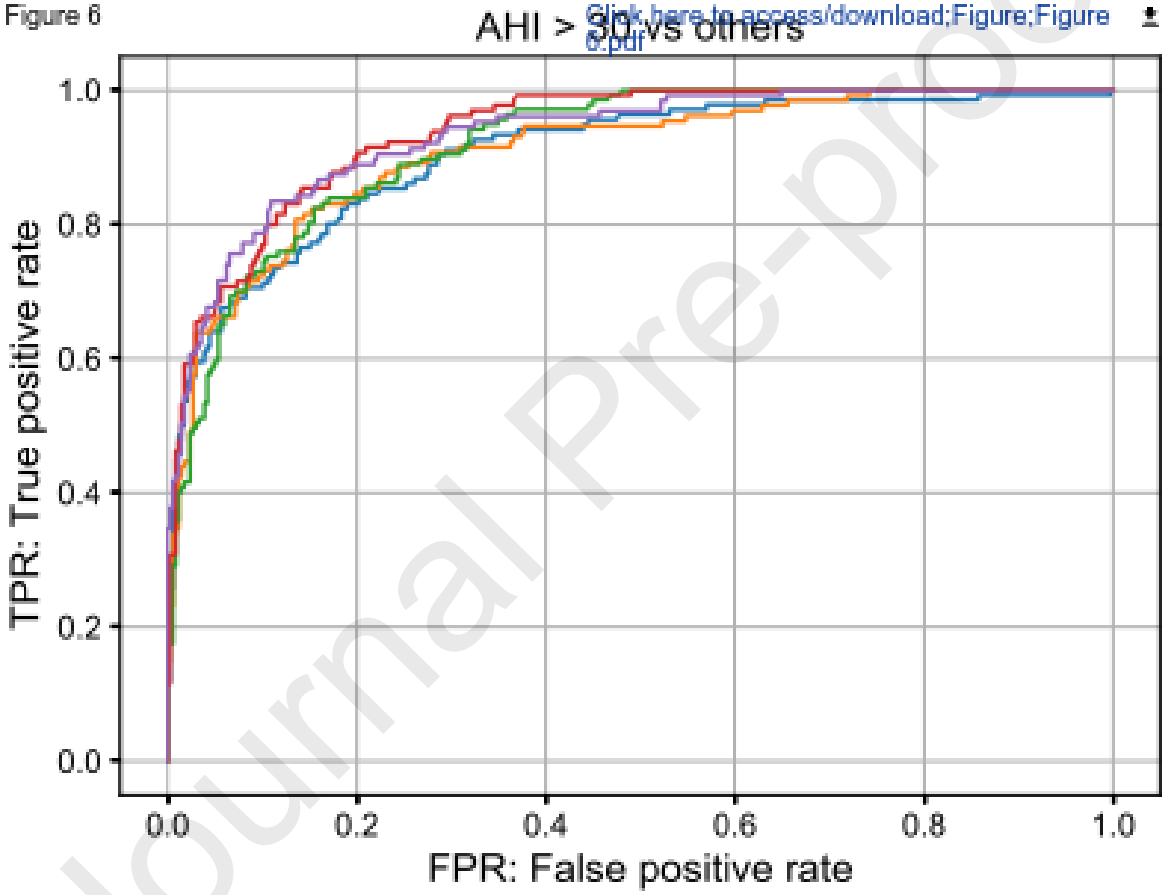


Figure 7

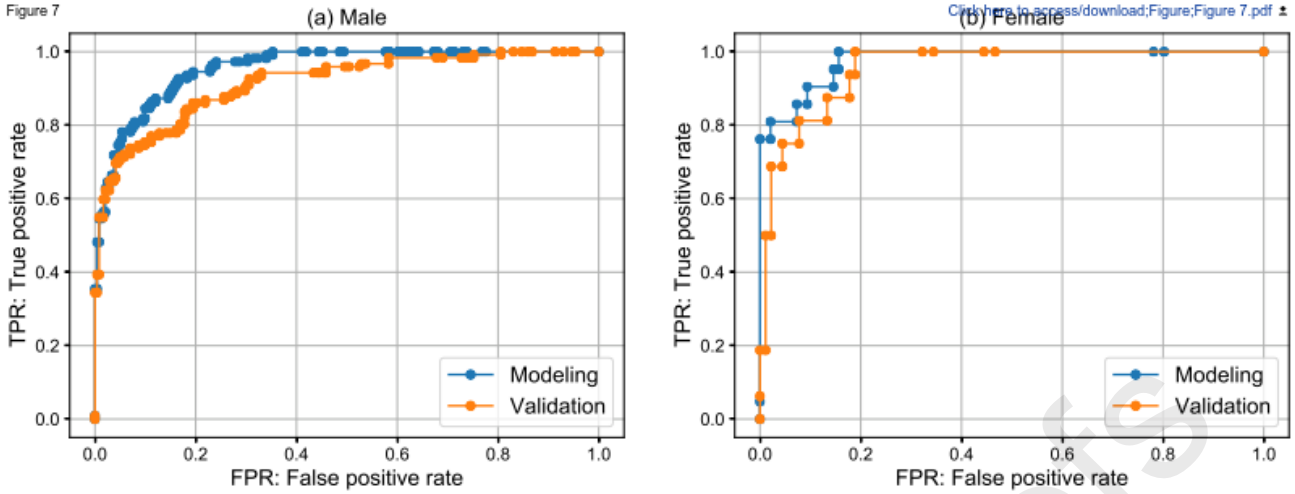


Figure 8