의학박사 학위논문

Development of sleep apnea prediction models from smartphone-recorded sleep breathing sound

스마트폰에 녹음된 수면 중 호흡음을 이용한 수면 무호흡의 예측 모델의 개발

2023 년 8 월

서울대학교 대학원
의학과 이비인후과학전공
조 성 우

Development of sleep apnea prediction models from smartphone-recorded sleep breathing sound

지도 교수 김 정 훈

이 논문을 의학박사 학위논문으로 제출함

2023 년 4 월

서울대학교 대학원

의학과 이비인후과학전공

조성우

조성우의 의학박사 학위논문을 인준함

2023년 7월

위 원 장 _____ (인)

부위원장 _____ (인)

위 원 _____ (인)

위 원 _____ (인)

위 원 _____ (인)

`

Abstract

Development of sleep apnea prediction models from smartphone-recorded sleep breathing sound

Sung-Woo Cho

Otorhinolaryngology

The Graduate School, Seoul National University

Introduction: Breathing sounds during sleep is an important characteristic feature of obstructive sleep apnea (OSA) and has been regarded as a potential biomarker. Breathing sounds during sleep can be easily recorded using a microphone, which is usually found in most smartphone devices. Therefore, it may be easy to implement as an evaluation tool for pre-screening purposes. The purpose of this study is to predict OSA using smartphone recorded sounds and identify optimal settings with regard to noise processing and sound feature selection.

Material and methods: A cross-sectional study was performed. Patients who visited the sleep center of a tertiary hospital for snoring or sleep apnea from August 2015 to August 2019 were enrolled. Audio recordings during sleep were performed using a smartphone during routine full-night in-lab polysomnography. A total of 423 patients were analyzed. Data were split into training (60%, n = 256) and test datasets (40%, n = 167). Using a random forest algorithm, binary classifications were separately conducted

`

for three different threshold criteria according to an apnea hypopnea index (AHI) threshold of 5, 15, or 30 events/h. Four regression models were created according to noise reduction and feature selection from the input sound to predict actual AHI; 1) noise reduction without feature selection, 2) noise reduction with feature selection, 3) without noise reduction and feature selection, and 4) feature selection without noise reduction. Clinical and polysomnographic parameters which may have affected errors were assessed.

Results: Accuracies were 88.16%, 82.25% and 81.66% and the areas under curve were 0.9, 0.89, and 0.9 for an AHI threshold of 5, 15, and 30 events/h, respectively. In the regression analysis, using recorded sounds that had not been denoised and had only selected attributes resulted in the highest correlation coefficient (r = 0.784, 95% confidence interval(CI): 0.689~0.879). AHI (beta = 0.329, 95% CI : 0.242~0.415) and sleep efficiency (beta = -0.197, 95% CI : -0.348~-0.046) were found to be related to estimation error.

Conclusions: Recorded sleep breathing sound using smartphones provides a reasonable prediction of OSA. Future research should focus on real life recordings using various smartphone devices.

Key words: Obstructive sleep apnea; sleep breathing sound; smartphone; prediction

Student number: 2018-39428

`

CONTENTS

`

`

List of figures and tables

`

`

List of abbreviations and symbols

AHI	Apnea hypopnea index
AUC	Area under curve
BMI	Body mass index
CI	Confidence interval
IQR	Interquartile range
NPV	Negative predictive value
OSA	Obstructive sleep apnea
PPV	Positive predictive value
SHAP	SHapley Additive exPlanation
SVM	Support vector machine
SMOTE	Synthetic minority over-sampling technique balancing method

`

CHAPTER 1

Prediction of OSA from sleep breathing sound: proof of concept study

`

Introduction

Obstructive sleep apnea (OSA) is a common disorder characterized by repeated upper airway closure during sleep causing intermittent hypoxia and arousal during sleep. It is known to be an independent risk factor several important cardiovascular and cerebrovascular disease and also results in a poor sleep quality and overall impaired quality of life [1-3]. The diagnosis of OSA is based upon the polysomnography which is usually executed in an in-lab setting with multi-channel monitoring in attended condition. Although polysomnography is a current golden standard, it is costly and time consuming. Furthermore, because of a limited number of centers that has this facility, many patients still remain undiagnosed.

Treatment of OSA includes continuous positive airway pressure (CPAP) which is widely used as the first line treatment. However, due to decreasing long-term patient adherences, there are other treatment options such as surgery, life style modification, and mandibular advancement device. During or after treatment, it is important to re-evaluate the patient status. Due to aforementioned limitations, follow up testing is also quite difficult. In short, the current standard diagnostic testing, polysomnography, is not always suitable for screening and follow up..

Breathing sound during sleep or snoring is an important characteristic feature of OSA and has been regarded as one of potential biomarkers for OSA. According to a pooled analysis, overall sensitivity in the diagnosis of OSA from the breathing sound during sleep is 0.88 and specificity is 0.81 [4].

Using breathing sound during sleep can be easily recorded by using a microphone which is mostly mounted in the smartphone these days and

`

therefore developing simple algorithm by using breathing sound during sleep for prescreening or follow up during or after treatment may eventually fulfill the unmet needs. Our previous study had demonstrated the possibility of prescreening of OSA by using binary classifier from the respiratory sounds during sleep[5-7]. Especially, we have simplified our study approach by using whole sound data during sleep and unbiased feature selection[6]. In this study, we tried to predict apnea hypopnea index (AHI) from breathing sound generated during sleep.

Methods

Study participants

Patients who visited the sleep center of a tertiary hospital due to snoring or sleep apnea were enrolled in this study. All of them underwent attended, in-laboratory, full-night polysomnography. For all patients, audio recordings were concurrently performed with an air-conduction microphone during polysomnography (**Figure 1**).

Sound processing

Analyses for predicting AHI were performed included all sleep stages from sleep onset till offset. The sound processing was conducted as previously described[6]. In short, Noise reduction preprocessing was followed by data segmentation into 5 second windows and audio features representing a variety of temporal and spectral characteristics of audio signal were extracted. These procedures were done with the jAudio, a Java-based audio feature extraction software. In total, 508 features were extracted from each patient. (**Table 1**)

`

Estimation of apnea-hypopnea index (AHI)

Estimation of AHI from was performed by regression analyses with the following methods: Gaussian process, support vector machine, random forest, and simple linear regression. The estimation of AHI models were constructed using 508 sound features and measured AHI (mAHI) as factors and covariates and estimated AHI (eAHI) as output. Validation was conducted with 10-fold cross-validation. In brief, enrolled patients were randomly divided into 10 equal-sized subgroups. Of the 10 subgroups, a single subgroup was retained for validation of the prediction model; the remaining nine subgroups were used for training. The cross-validation process was then repeated 10 times (10 folds), with each of the 10 subgroups used once for validation. This gives 10 evaluation results, which are averaged. Then learning algorithm was then invoked a final (11th) time on the entire dataset to obtain the final model [6, 8]. Estimation process is

summarized in **Figure 2.**

Evaluation of prediction model performances

Model performances were evaluated by correlation coefficient between mAHI and eAHI. Several errors including mean absolute error, root mean squared error, relative absolute error, and root relative squared error were also assessed. Mean absolute error and root mean squared error simply measures the average differences so that they are in the same scale of the measured variable. In relative absolute error and root relative squared error, average differences are divided by the variation to they have a scale from 0 to 1. Other performance evaluation metrics included bias, precision, and accuracy).Bias was defined as the median of the difference between measured AHI (mAHI) and estimated AHI (eAHI). Precision was defined as the interquartile range

`

(IQR) of the difference [9, 10]. Accuracy was the proportion of participants whose eAHI was not deviated more than 50% from mAHI (P_{50}). The 95% confidence intervals were calculated by the bootstrap method (2,000 boot straps) [11].

Statistical analyses

Significant testing of the differences in performance was done by Wilcoxon signed rank test (for model comparison) or Kruskal-Wallis test followed by Mann-Whiteny test for post hoc comparison (for subgroup comparison). McNemar test was used for model comparison and linear by linear association was used for accuracy in subgroup comparison. Other clinical and demographic factors were compared by 1-way analysis of variance (ANOVA). Statistical analyses were performed by using SPSS ver. 22.0 (IBM Corp., Armonk, NY, USA). Machine learning was performed with a free software, Weka [8]. Statistical significance level was $P < 0.05$. This study was approved by the Institutional Review Board of Seoul National University Bundang Hospital (IRB No. B-1404/248-109).

Results

Clinical and polysomnographic characteristics of patients

A total of 116 patients were analyzed. Patients were grouped according to their mAHI into normal (mAHI < 5; n = 28), mild (5 ≤ mAHI < 15; n = 28), moderate (15 ≤ mAHI < 30; n = 30) and severe (30 ≤ mAHI; n = 30). The mean age and mAHI of all patients was 50.4 ± 16.7 years and 23.0 ± 24.0/hr, respectively. Distribution of mAHI among patients is described in **Figure 3.**

`

The mean total sleep time was 369.7 ± 104.8 minutes. Body mass index, and male to female ratio was significantly different among groups ($P < 0.05$). Apnea index, hypopnea index, mean duration of apnea, mean duration of hypopnea, and snoring time were also significantly different according to OSA severity ($P < 0.05$, **Table 2**)

Performance of AHI prediction

AHI prediction performance measurements are summarized in Table 2. The correlation coefficient between mAHI and eAHI was the highest at 0.83 in the random forest model. The least mean absolute error, root mean squared error, relative absolute error, root relative squared error were also the least at 9.64/hr, 13.72/hr, 0.52, and 0.57 respectively in random forest. Other models resulted somewhat lower but similar performance with correlation coefficient ranging from 0.74-0.79. Regression plot and Bland-Altman plot based on the eAHI estimated using random forest are presented in **Figure 4.** Bland-Altman plot showed that the AHI mean difference between mAHI and eAHI was about 0.46/hr and that eAHI tended to be underscored as the severity of OSA increases. Performances in prediction of apnea index and hypopnea index were also evaluated. Estimation of apnea index (correlation coefficient = 0.78 - 0.83) showed overall better performance regardless of estimation models compared to that of hypopnea index (correlation coefficient = 0.15 - 0.47) (**Table 3**).

Bias, precision, and accuracy: for subgroup analysis of performance

In order to evaluate the estimation model according to disease severity, subgroup analysis by mAHI was performed by using bias, precision, and accuracy (**Table 4**). There was no significant difference in overall bias (median

`

difference between mAHI and eAHI) among estimation models ($P > 0.05$), however the overall bias tended to be smallest in support vector machine (0.25/hr), followed by Gaussian process (0.79/hr), and simple linear regression (1.30/hr). Random forest resulted in highest overall bias (3.41/hr). In subgroup analysis, a significant difference in bias between models was found only in the normal mAHI group in comparison between support vector machine and random forest ($P=0.031$). Regardless of estimation models, bias in severe OSA group was significantly greater than other groups ($P < 0.05$). The best (which means the lowest) overall precision (IQR of the difference) had been achieved in random forest model (12.07/hr) followed by simple linear regression (16.46/hr), support vector machine (16.49/hr) and Gaussian process (19.36/hr). The severe group had the worst precision compared to the other severity groups regardless of estimation models. Accuracy which indicates the percentage of estimates that differed from the measured AHI by less than 50%, was similar for all models across all mAHI subgroups ($P > 0.05$). However, all models resulted in significant differences in accuracy according to OSA severity with tendency of higher accuracy in moderate to severe groups compared to mild to normal ($P < 0.05$).

Most correlating sound feature

The prediction using simple linear regression was comparable in all performance metrics, compared to other methods and revealed that derivative of area methods of moments overall standard deviation (feature 188) showed the highest correlation with mAHI. The regression formula was proposed as eAHI = 24.92 X Feature188 - 27.83. This led to overall coefficient of 0.79, mean absolute error of 10.75/hr, and root mean squared error of 14.76/hr.

`

Discussion

The current study validated the proof of concept that from the sound data during sleep, we may estimate the actual severity of OSA as represented with eAHI. Therefore, our algorithm entails certain potentials to be used for prescreening of OSA and also for repeated follow up studies to estimate the therapeutic efficacy of treatment. For example, during lifestyle modification including weight reduction and exercise, we can monitor its effect and consult the patients based on the eAHI because repeated polysomnography is almost impossible in real practice settings.

There had been some studies that used snoring sound to diagnose OSA, by using several different methodologies. However in these studies, manually labeled snoring sound data was required, and hypothesis driven approaches with mathematical modelling of the snoring sound to estimate the apnea event were necessary [12-15]. In contrast to other studies, we have simplified our methods by collecting breathing sound during sleep in all stages from sleep onset to sleep offset. We also tried to extract unbiased sound features as much as we can without any presumptions. Prediction accuracy may have been lowered due to abundance relatively unimportant data, however we tried to overcome this limitation by using machine learning.

In our model, overall performance had been greatest in random forest method which is an ensemble learning method. In ensemble learning method, different models are combined to generate better result [16]. However, our results also showed that in random forest, the bias was somewhat higher compared to other methods meaning that developing an optimal model with ensemble method is rather difficult to superb other regression models. In our study, Gaussian

`

process and support vector machine which are both kernel-based methods showed somewhat smaller bias compared to random forest method. However, from precision point of view, random forest showed better performance. In Gaussian process and support vector machine, a single good model might be constructed from theoretical data, however, this single model may have a high variance. Therefore, these models would have relatively low bias relative to the variance. On the contrary, combining several models as in random forest, the overall variance may be decreased[16].

Accuracy had been higher in moderate and severe groups. Among AHI subgroups, moderate OSA group seems to be the most accurately predicted group considering the low bias and high accuracy and this is also consistent with our previous finding which evaluated a binary classification of OSA[6]. Accuracy incorporates both bias and precision and among evaluation metrics for accuracy, we used P_{50} (deviated more than 50% from mAHI) which is somewhat arbitrary. Therefore this needs further validation. This value is also a relative measure that accuracy varies according to the level of AHI and does not have consistent meaning across the whole range. However other evaluation metrics for accuracy include mean squared error or its square root, and these too are measured by the log scale that they also have the same drawback[10].Another important finding in our study is that prediction performance in severe OSA had been worse regardless of learning models. In these groups, there was an underestimation of AHI. There may be several reasons for this. Firstly, although patients in our cohort are rather equally distributed according to AHI severity, the actual distribution of the AHI values is somewhat skewed with severe AHI being less frequent. Therefore, learning from the input data may not have been sufficient enough in cases of severe OSA. Secondly, the duration of apnea in severe OSA group was higher than

`

other groups and especially, when compared to normal or mild OSA groups, the difference is significant. In cases of flow limitation, airway starts to vibrate therefore generates a sound which is typically known as snoring. However, in cases of apnea, when the respiration ceases with no airflow in the upper respiratory track, breathing sound becomes quiescent. An abrupt breathing sound occurs when apnea event has ended and respiratory related arousal begins. Therefore if apnea duration is longer, sound features may be less recorded with underestimation of feature derivatives.

Prediction performance of hypopnea index turned out to be poor. However, in our study, we tried to predict AHI which is a summation of apnea index and hypopnea index, since the clinical significance of hypopnea is also important as well. The consequences of hypopnea is known to be similar to that of apnea regarding oxygen desaturation, and EEG arousal, and increase in heart rate[17]. Therefore, clinically, it is important to predict AHI rather than apnea index or hypopnea index seperatively. However, hypopnea index and its proportion among AHI are significantly different of OSA subgroups, and this may partially explain the performance differences among OSA subgroups.

Interestingly, simple linear regression analysis had resulted in fairly comparable outcome. In simple linear regression, correlation between 508 features and AHI are performed and among them feature 188 turned out to be the most significant correlated feature. This is a derivative of area methods of moments overall standard deviation. In our study, derivatives of each feature were calculated to observe temporal changes and further, standard deviations of derivatives were also calculated. Methods of moments describe numeric quantities at some distance from a reference point or axis. Therefore, area methods of moments describes the shape of spectrogram [18]. The performances from simple regression analysis with feature 188 were

`

comparable to other machine learning methods suggesting a linear correlation between AHI and sound features.

There are several hurdles to overcome for our algorithm in order to be applied in a real world. First of all, overall performance should be further increased. As was mentioned above, more learning from severe OSA patients (unbiased sampling) is necessary to improve the performance. Second, sound features are known to be different according to anthropometric measurements [19, 20], however our algorithm did not consider such parameters. And finally, since our ultimate goal is to apply our algorithm to a smart phone via mobile apps, sound data from smart phone recording should be validated as the same manner. Technical issues also exist. Data processing system which incorporates denoising, feature selection, and machine learning should be established in a mobile phone setting. Nonetheless, this is a proof of concept and further data acquisition and technical development is on the way.

Conclusion

It seems that AHI could be predicted using breathing sound generated during sleep with a good performance. With more machine learning from the sound features and measured AHI from additional patients and further validation, our prediction model may be useful not only for pre-screening but also for a follow up after treatment in patients with OSA.

`

CHAPTER 2

Prediction of Sleep Apnea from Smartphone Recorded Sleep Breathing Sounds

`

Introduction

Obstructive sleep apnea (OSA) is a common disorder that is widely known to be associated with decreased quality of life and increased incidence of neurovascular and cardiovascular diseases[21]. The diagnostic gold standard method for OSA is attended full-night in-lab polysomnography (PSG) that involves recording numerous physiologic signals that are manually scored by certified sleep technicians or physicians. Therefore, in-lab PSG is expensive and accessibility to sleep facility is not always easy. Considering the high prevalence of OSA, performing full night in-lab polysomnography may not be practical for all patients. Another major drawback is that the sleep environment during PSG is not exactly the same as in real life and there is considerable night-to-night variation[22-24]. Consequently, portable home sleep apnea test devices were developed and have been used in selected cases[25] as screening tools and for patient monitoring during and after treatment. However, in order to use these devices, patients must obtain a phiscian's prescription, and the high cost limits access to the general public

Meanwhile, sleep breathing sounds, which include snoring, are biomarkers that may represent OSA[26, 27]. In our previous study, we focused on the prediction of OSA from sound data collected during sleep using a conventional camcorder microphone[28, 29]. The advantage of utilizing recorded sleep breathing sounds in the detection of OSA is that it can be repeatedly tested. In addition, the widespread use of personal smartphones, which have a microphone recorder, makes it easy to implement.

This study focused on the prediction of OSA from sleep breathing sounds

`

recorded from a conventional smartphone. In our previous work, after sound was recorded, data was processed and features were extracted. However, in a smartphone environment, data processing and feature extraction need to be minimized.

The primary goal of this study was to assess whether smartphone recorded breathing sounds could be used to predict OSA. Our secondary goal was to evaluate whether minimizing features and sound processing could affect prediction performance.

Methods

Study participants and full-night PSG

A cross-sectional study was performed. From September 2015 to September 2019, patients who visited the department of Otorhinolaryngology Head and Neck Surgery Sleep Clinic of a tertiary hospital due to snoring or sleep apnea were included in this study. All patients underwent an attended in-laboratory full-night PSG (Embla N 7000, Reykjavik, Iceland). Written informed consent was obtained from each participant and the study complied with the Declaration of Helsinki. This study was approved by the Institutional Review Board of Seoul National University Bundang Hospital (Seongnam, Korea; IRB No. B-1404/248-109).

Apnea was defined as cessation of airflow for at least 10 s; hypopnea was defined as a >50% decrease in airflow for at least 10 s or a moderate reduction in airflow for at least 10 s associated with arousals or oxygen desaturation (<4%)[30]. The apnea hypopnea index (AHI) was defined as the total number of apnea and hypopnea events per sleep hour. Patients with analysis time periods (time spent in bed) that lasted at least 4 h were included.

`

Sound recording and processing

For all patients, audio recordings were carried out using an LG G3® smartphone (LG, Seoul, Korea) during full-night PSG. The smartphone was consistently placed at the corner of a shelf and was approximately 1 m from the patients' head (**Figure 5**). Sleep breathing sounds were recorded for all sleep stages, from sleep onset to offset.

Sound analyses began with converting audio-files into wav file formats. Audio data were discarded for the initial 30 min from initiation of recording considering the median time of sleep latency as 9.0 min (interquartile range: 11.5 min). Analysis was ceased once the analysis time exceeded 6 h. First, noise filtering was performed with a spectral subtraction using Audicity®[31], followed by data segmentation into 5-s windows; thereafter, sound features were extracted from each windowed signal. These procedures were performed using jAudio, a Java-based audio extraction program[32]. We attempted to extract all sound features provided by the software. Means, standard deviation, derivatives of means, and derivatives of standard deviations of each feature were then calculated. Finally, 508 features, representing a variety of temporal and spectral characteristics of the audio signal of sleep breathing sounds, were extracted from each patient (**Table 1**). The process was repeated for the recorded sounds without noise filtering.

Dataset portioning and machine learning

Partitioning of the dataset into training and test dataset and machine learning was performed using the free software WEKA[33]. The data of the patients were first randomly sorted and divided into training (60% of patients) and test (40% remaining patients) datasets by using Randomize and

`

RemovePercentage filter. Each patient yielded two datasets acquired from two different conditions (with and without noise filtering).

First, binary classifications were conducted for three different threshold criteria of AHI: 5, 15, or 30 events/h. The synthetic minority over-sampling technique balancing method (SMOTE) was used to introduce a balance in the training dataset[34]. A binary classification of AHI was carried out using a random forest method, as this demonstrated the best performance according to a previous study[29]. A ten-fold cross validation procedure was applied. The final models were then assessed in an independent test dataset.

The estimation of AHI was performed by creating regression models, also using the random forest method in the training dataset. Regression models were developed based upon noise reduction and feature selection. Feature selection involved selection of "n" number of relevant features from the original 508 sound features, and this was done using correlation-based Feature Subset Evaluation in WEKA[35]. In total, four regression models were developed as follows: noise reduction only (model 1), noise reduction with feature selection (model 2), naïve recorded sound only (without noise reduction) (model 3), and naïve recorded sound with feature selection (model 4).

Model performance measures, including accuracy, kappa index, sensitivity, specificity, F-1 score, area under the precision recall curve, and the area under the curve (AUC) of the receiver operating characteristic curve, were computed for the binary classification. AHI estimation was evaluated; mean absolute error and root mean squared error were used for error metrics. Finally SHapley Additive exPlanation (SHAP) values were calculated to provide attribution values for each feature in the best prediction model by using Python (version 3.5). The SHAP value evaluates the significance of the

`

output resulting from the inclusion of a particular feature for all other feature combinations.[36]

Statistical analyses

After confirmation of non-normal distribution, a Mann-Whitney U test was used to compare the mean scores of clinical variables of the training and test datasets. A chi-squared test was used for categorical variables. A pearson correlation and multivariate linear regression models were carried out to examine the associations between error and clinical parameters. Age, sex, body mass index (BMI), AHI, and sleep efficiency were used as input variables in the models. Statistical analyses were performed using SPSS version 22.0 (IBM Corporation, Armonk, NY, USA). Continuous parametric variables are presented as means ± standard deviations. A p-value < 0.05 was considered statistically significant.

Results

Patient

During the study period, 760 patients underwent polysomnography. After excluding patients under age 18 years and those who refused to enroll in this study, a total of 423 patients were included. Patients were grouped according to OSA severity: normal (N = 43, mean AHI = 2.2±1.5 events/h), mild (N = 80, mean AHI = 9.4±2.8 events/h), moderate (N = 109, mean AHI = 22.0 ± 4.2 events/h), and severe (N = 191, mean AHI = 55.1 ± 17.3 events/h). **Table 5** describes the main characteristics of the analyzed patients. There were no significant differences in demographic and polysomnographic parameters between the training and test datasets.

`

Performance of binary classifiers for OSA

When the AHI threshold for binary classification of OSA was 5, 15, and 30 events/h, the accuracy of the OSA prediction was 88.17% (Cohen's kappa coefficient (κ) = 0.46), 82.84 % (κ = 0.59) and 81.65% (κ = 0.63), respectively. Sensitivity was 90.79%, 87.29%, and 82.95%, respectively. Positive predictive value (PPV) was 98.83%, 89.29%, and 82.02% for AHI thresholds of 5, 15, and 30 events/h respectively. Accuracy, sensitivity, PPV, F-1 score and area under precision recall curve tended to decrease as the threshold for OSA was increased. However, the kappa value, specificity and negative predictive value (NPV) increased along with the threshold: specificity was 64.71%, 70.59%, and 80.25%, respectively, for AHI thresholds of 5,15, and 30 events/h, and NPV was 44.0%, 70.59%, and 81.25%, respectively. AUC values were 0.902, 0.885, and 0.896, respectively, for AHI thresholds of 5,15, and 30 events/h (**Table 6**).

AHI estimation and effect of denoising and attribute selection

AHI estimation was carried out using four regression models. Selected attributes (model 2 and 4) and their main characteristics are summarized in Supplementary **Table 7 and 8**.

All models resulted in similar results with a correlation coefficient ranging between 0.77 and 0.78. Using sleep breathing sounds without any denoising and attribute selection (model 4) yielded the best estimation performance. The correlation coefficient was 0.784 and the root mean squared error was 14.73 events/h. Other metrics are described in **Table 9**. The correlation plot and Bland-Altman plots acquired from model 4 (feature selection without

`

noise reduction) are described in **Figure 6**. The Bland-Altman plot revealed that the mean difference between measured AHI and predicted AHI was approximately 0.23 (95% confidence interval, -28.73, 29.18) events/h and error tended to increase as the mean AHI increased (error = -7.503 + 0.242 (95%CI: 0.137~0.347)*mean AHI, R^2=0.110) pointing to the underestimation of AHI as the OSA severity increases. SHAP values calculated from the model 4 are described in **Figure 7.** The highest SHAP value of 15.45 was derived from the feature named "Derivative of Relative Difference Function Overall Standard Deviation"

Factors that contribute to error

A correlation analysis was carried out on several clinical parameters (AHI, sleep efficiency, BMI, and age) and error in AHI estimation (measured AHI–estimated AHI based on model 4). Factors that were significantly associated with errors were AHI (r = 0.594, 95% CI : 0.481~0.717), sleep efficiency (r = -0.165, 95% CI : -0.316~-0.015), BMI (r = 0.359, 95% CI : 0.217~0.502), and age (r = 0.164, 95% CI : 0.013~0.314) (**Figure 8**). A multivariate analysis using a linear regression model revealed that AHI (beta = 0.329, 95% CI : 0.242~0.415) and sleep efficiency (beta = -0.197, 95% CI : -0.348~-0.046) was significantly associated with error in estimation (**Table 10**).

Discussion

This study validated the utility of breathing sounds recorded during sleep using a smartphone microphone for OSA prediction. To the best of our

`

knowledge, this is the largest cohort of patients who simultaneously had their sleep breathing sounds recorded by a smartphone during a standard full-night in-laboratory PSG.

Our previous studies included a small number of patients and sleep breathing sounds recorded on a low quality microphone[28, 29]. We had conducted a 10-fold cross validation without validating through an independent test set that there would be a chance of overfitting. This study lowered the possibility of overfitting by separating the entire dataset into training, validation, and test datasets[37]. However, the patient distribution is somewhat different from that in our previous study. In our previous study, the number of patients was similar in each OSA severity group; however, in the current study, the number of patients tended to increase as the OSA severity increased. Imbalanced data could have affected accuracy. Instead, to overcome the imbalanced patient distribution, we utilized SMOTE in the training set to augment the learning from the minority class. The concept of **SMOTE** is the generation of synthetic data between each sample of the minority class and its "**k**" nearest neighbors. Therefore, the training dataset was more balanced than before, and the problem of overfitting during model buildup was alleviated with SMOTE[34]. Consequently, the OSA prediction performance was comparable to that in our previous study. The prediction accuracy of OSA ranged from 81.65–88.17%.

The presence of OSA with a cut off value of 5 to 30 events/h could be predicted with a sensitivity of 82.95 to 90.79% and a specificity of 64.71 to 80.25% using breath sounds recorded with a smartphone during sleep. However, even though we balanced our data in the training set using SMOTE, the binary classification performance seemed to depend on the cut off threshold. As the threshold was increased, sensitivity was lowered while

`

specificity was increased. The typical trade-off between sensitivity and specificity was observed as the threshold value changed: decreasing sensitivity and increasing specificity as the threshold increased. This has also been observed in other home sleep apnea test devices[38]. In general, the prevalence of OSA is 28.4% in Korea and 33.2% for USA for AHI≥5 events/h[39] which is lower than that of the current study. Therefore, predictive values would be adjusted when used for general population for screening purposes; PPV (82.02~95.83%) would be further decreased as NPV (44.0~81.25%) increased.

In our study, several metrics including accuracy, kappa index, and receiver operation curves were used to assess the predicve performance based on various AHI threholds. Although, accuracy, F-1 score, area under precision recall curve and AUC values were highest for AHI≥5 events/h, the kappa index which measures the proportion of correctly instances after accounting for the probability of chance agreement was highest when the threshold for AHI were set up to ≥30 events/h. In our previous study, where there had been an even distribution of the number of patients in each OSA severity, the highest kappa index was shown for AHI≥15 events/h[28]. The ratio between two classes above or below this level was 1:1 suggesting a balanced distribution for both training and testing. In the current study, although the input sound data was technically different, the majority of patients in the test data set were in the severe OSA category (47.9%). Therefore, when the cut off value had been set up at 30 events/h, the proportions of the classes below or above this level seem to be the most balanced. Data distribution in the test dataset impacts on the model performance[40] even after artificially balancing the data with SMOTE.

Overall, our result is comparable to portable home sleep test devices

`

(sensitivity :0.88, specificity 0.88, AUC: 0.888 for AHI threshold 15 events/hr) [41] and one recenetly published paper from wearable device based on reflective photoplethysmography (sensitivity:0.73, specificity: 0.81, AUC: 0.86)[42]. Unlike wearable or home sleep test devices, our method is easier to implement and repeatable because breathing sounds can be easily obtained from a smartphone. Therefore, our method has potential for pre-screening for OSA. However, even if the result is negative, for high risk patients, such as adults presenting signs and symptoms (excessive daytime sleepiness, obeses, habitual snoring, or diagnosed hypertension) of OSA[43], standard in-lab PSG or at least a home sleep apnea test is should be recommended.

In our previous model based on a ceiling microphone in a PSG laboratory, input sound data underwent noise reduction followed by feature extraction[28, 29]. The process of noise reduction by spectral subtraction filtering and the extraction of all 508 sound features from sleep recorded sound data demand additional resources. As our ultimate goal was to predict or prescreen OSA using smartphones, it would beneficial for the process to be simple. Interestingly, using sound data that had not been processed and had only selected features yielded a better prediction performance compared to that of sound data that had been denoised and used all 508 sound features. Model 4, which was based on sound data that had not been denoised and contained selected features, resulted in the best prediction performance. Attribute selection removes irrelevant features that actually decrease the prediction performance by introducing additional noise. This is known to decrease overfitting, improve accuracy, and reduce the training time[35, 44]. Most smartphones use adaptive noise cancellation by an additional microphone[45]. Adding a second microphone to sample the noise of the

`

acoustic environment allows for that signal to be subtracted from the sound recorded using the original microphone[46]. Recorded sounds on smartphones have already been filtered once; therefore, noise reduction by spectral subtraction may not be necessary.

Conclusion

Previous proof of concept study validated the usefulness of using sleep breathing sound to estimate OSA. This study suggests that recorded sleep breathing sound using smartphones provides a reasonable prediction of OSA. Future research should focus on real life recordings using various smartphone devices.

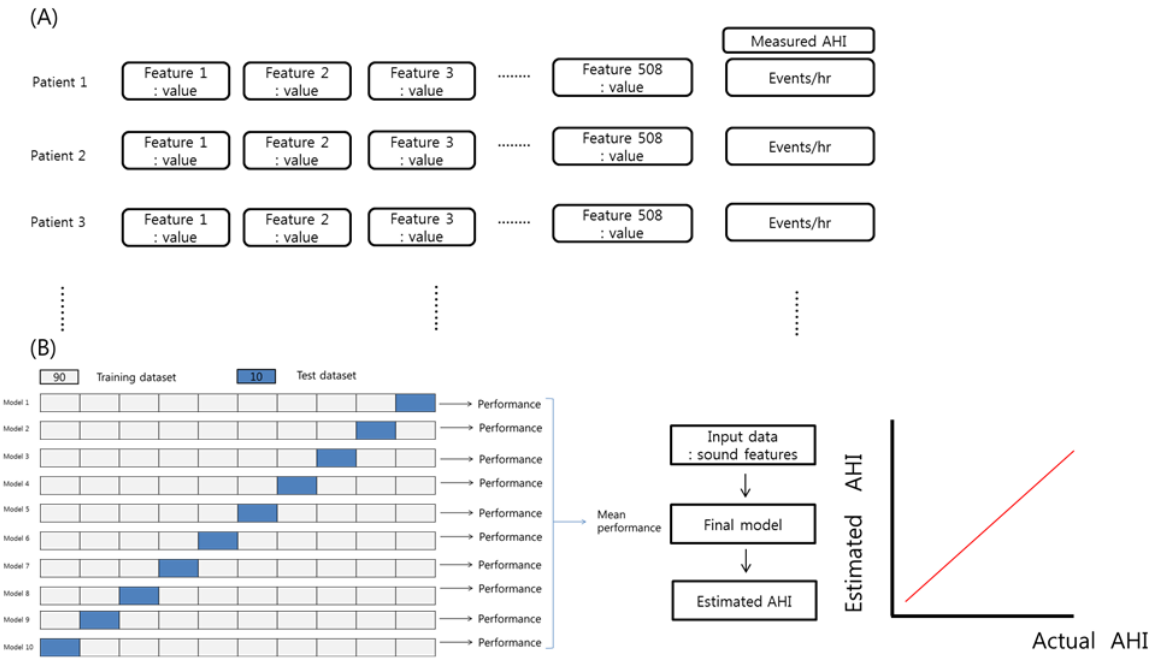
`

Figure 1.

A bed for polysomnography and a microphone (inset) on the ceiling.

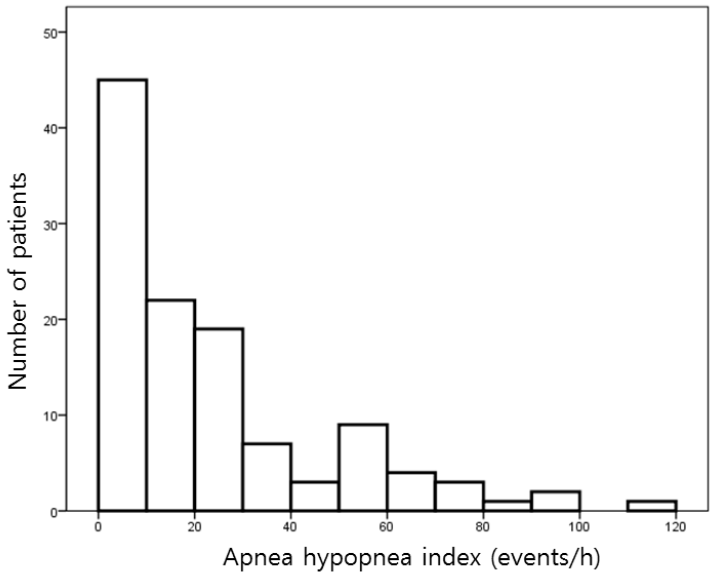
`

Figure 2. Machine learning process for apnea-hypopnea index (AHI) prediction. (A) In total, 508 sound features were extracted during full-night polysomnography. (B) After a 10-fold cross-validation process, the learning algorithm was invoked on the entire data set to obtain the final model.



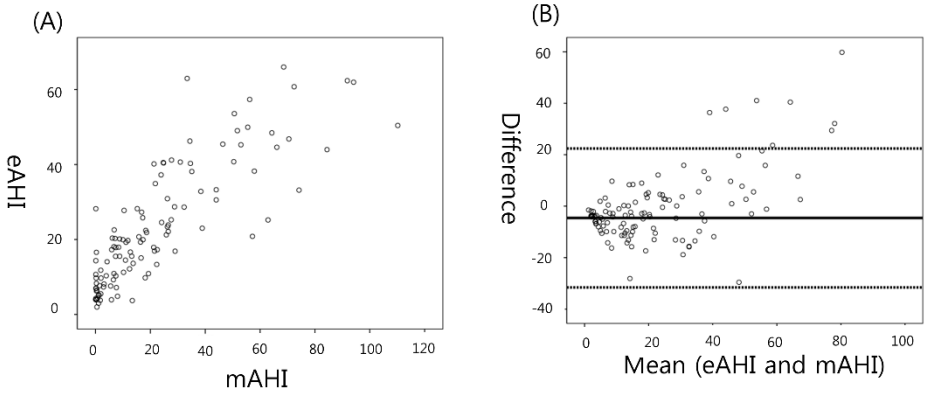
`

Figure 3. Histogram of apnea-hypopnea index (AHI) measured through polysomnography. Distribution of the AHI values was skewed left, with severe AHI being less prevalent.



`

Figure 4. Regression plot (A) and Bland-Altman plot (B) based on estimated apnea-hypopnea index (eAHI) using random forest. Correlation coefficient between eAHI and measured apnea-hypopnea index (mAHI) was 0.83. Both plots demonstrate eAHI tended to be underscored as the severity of obstructive sleep apnea (OSA) increased.



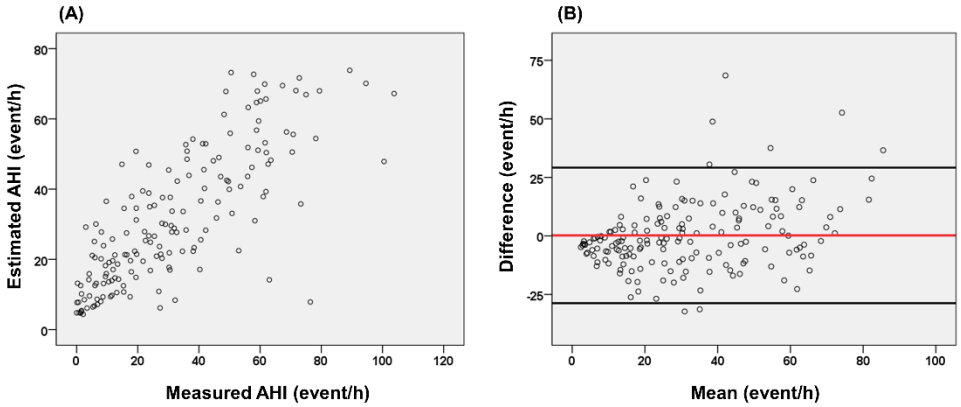
`

Figure 5. A bed for polysomnography and a smartphone on a shelf. The distance between the patient's head and smartphone is approximately 1 m.

`

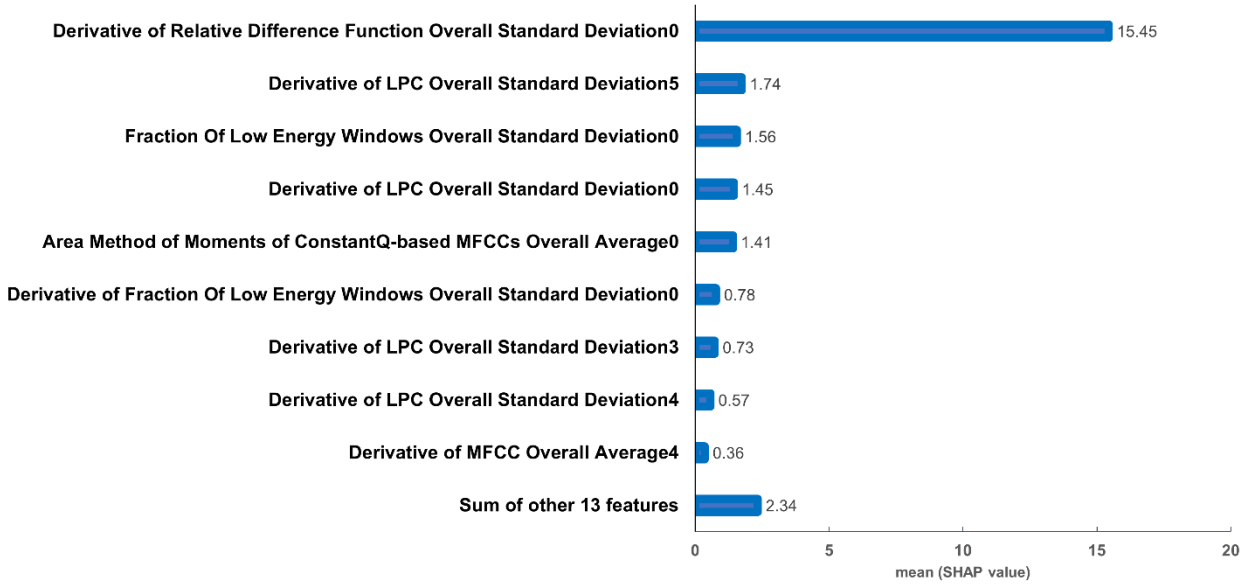
Figure 6. **Correlation plot (A) and Bland-Altman plot (B) of estimated AHI and measured AHI.** The estimation of AHI is based on model 4 which uses recorded sound without denoising followed by attribute selection. The correlation coefficient is 0.784 (A), and the error tended to increase as the mean AHI (mean of estimated AHI and measured AHI) increased (B) (error = -7.503 + 0.232*mean AHI, $R^2 = 0.110$, $p < 0.001$).

AHI: apnea hypopnea index



`

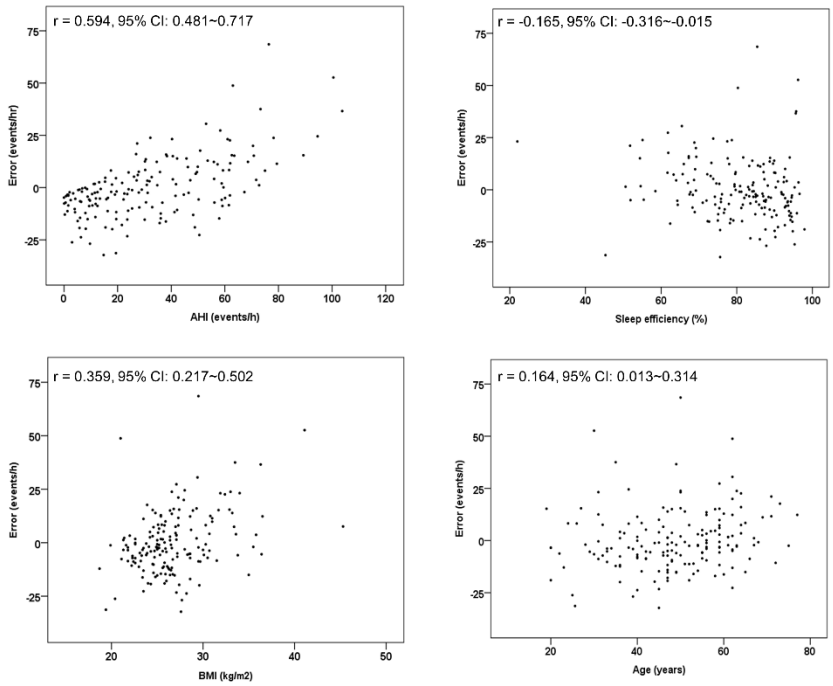
Figure 7. SHAP summary plot of the top nine features of model 4.



`

Figure 8. Correlation plot of error in AHI estimation (measured AHI – estimated AHI) and several clinical parameters

AHI; apnea hypopnea index



`

Table 1. Features extracted from respiratory sounds during sleep
A full-length audio data of each patient from sleep onset to sleep offset was segmented into a 5 sec window. With zero overlap, over 5 sec, 40000 samples were analyzed, and 127 audio features were calculated. Overall mean, standard deviation, derivatives of mean, and derivatives of standard deviation of audio features were then calculated. This led to generation of 508 audio features for each patient on entire sleep

Name of features	Number of derived features
Beat Histogram	172
Area Method of Moments	100
Mel Frequency Cepstrum Coefficient	52
Linear Predictive Coding	40
Area Method of Moments of Constant Q-based Mel Frequency Cepstrum Coefficients	20
Area Method of Moments of Log of Constant Q transform	20
Area Method of Moments of Mel Frequency Cepstrum Coefficients	20
Method of Moments	20
Beat Sum	4
Compactness	4
Fraction Of Low Energy Windows	4
Peak Based Spectral Smoothness	4
Relative Difference Function	4
Root Mean Square	4
Spectral Centroid	4
Spectral Flux	4
Spectral Rolloff Point	4
Spectral Variability	4
Strength Of Strongest Beat	4
Strongest Beat	4
Strongest Frequency Via Fast Fourier Transform Maximum	4
Strongest Frequency Via Spectral Centroid	4
Strongest Frequency Via Zero Crossings	4
Zero Crossings	4
The total number of features	**508**

`

`

Table 2. Baseline characteristics

Parameters	Normal (N=28)	Mild (N=28)	Moderate (N=30)	Severe (N=30)	P-value
AHI[1](events/h)	1.13±1.20	8.95±2.59	22.06±4.30	57.55±19.67	<0.001
BMI[2](kg/m^2)	23.15±3.94	24.68±3.23	26.93±3.16	27.31±4.19	<0.001
Male (%)	35.7	64.3	80.0	86.7	<0.001
Age (year)	43.21±20.58	54.00±14.58	53.93±13.30	48.63±18.84	0.057
Apnea index (events/h)	0.22±0.45	3.18±2.83	7.5±5.25	37.68±23.96	<0.001
Hypopnea index (events/h)	0.79±1.04	5.54±2.93	13.42±5.68	15.59±13.21	<0.001
Apnea index/Hypopnea index	0.53±1.76	1.26±2.25	0.89±1.18	7.78±11.41	<0.001
Mean apnea duration(sec)	11.42±13.30	18.91±8.97	22.88±7.00	25.29±7.30	<0.001

[1] AHI: apnea hypopnea index
[2] BMI: body mass index

`

Mean hypopnea duration(sec)	19.78±10.10	28.18±9.00	26.93±6.39	23.85±6.88	0.001
Snoring time (%)	12.67±12.65	25.23±21.90	29.92±23.78	20.77±14.95	0.006

`

Table 3. Estimation performances

Parameters	Method	Correlation coefficient	Mean absolute error (events/h)	Root mean squared error (events/h)	Relative absolute Error	Root relative squared error
Apnea hypopnea index	Gaussian Process	0.75	12.32	17.02	0.67	0.71
	SVM[1]	0.74	12.52	17.78	0.68	0.74
	Random Forest	0.83	9.64	13.72	0.52	0.57
	Simple linear	0.79	10.75	14.76	0.58	0.62
Apnea index	Gaussian Process	0.83	7.88	11.49	0.54	0.55
	SVM	0.83	8.31	11.79	0.57	0.57
	Random Forest	0.83	7.57	12.29	0.52	0.59
	Simple linear	0.78	9.46	12.98	0.65	0.63
Hypopnea index	Gaussian Process	0.15	8.21	11.91	1.14	1.25
	SVM	0.17	7.88	11.18	1.09	1.17
	Random Forest	0.38	6.32	8.82	0.87	0.92

`

	Simple linear	0.47	6.02	8.34	0.83	0.87

[1] SVM:Support vector machine

`

Table 4. Bias, precision and accuracy of each model

Variable	Overall	Normal	Mild	Moderate	Severe**
Bias = median difference (95% CI[1]) (events/h)					
Gaussian	0.79 (-1.18~3.96)	3.73 (-0.56~7.39)	4.46 (0.35~10.87)	0.28 (-3.74~5.22)	-11.29 (-18.76~-2.85)
SVM[2]	0.25 (-2.64~3.13)	4.87* (-0.89-7.31)	2.41 (-1.17~6.72)	0.97 (-4.69~5.56)	-10.57 (-21.70~-1.16)
Random forest	3.41 (-1.78~4.17)	5.96 (4.08~7.28)	5.47 (2.54~9.74)	1.54 (-2.73~4.81)	-10.18 (-17.77~-3.16)
Simple	1.30 (-1.79~4.87)	5.31 (0.33-8.76)	7.95 (1.22~9.27)	0.60 (-2.17~7.23)	-12.27 (-18.10~-4.10)
Precision = IQR[3] of the difference (95% CI) (events/h)					
Gaussian	19.36 (13.96~23.85)	12.34 (7.86~20.20)	12.74 (9.25~27.66)	14.74 (8.80~24.36)	31.42 (16.41~57.75)
SVM	16.49 (12.19~21.05)	14.67 (6.88~18.31)	13.43 (7.42~23.37)	16.71 (10.34~27.28)	28.82 (20.04~64.34)
Random forest	12.07 (9.47~16.42)	5.42 (2.70~8.70)	9.82 (6.35~12.64)	15.54 (7.69~18.72)	24.68 (14.43~37.53)
Simple	16.46 (13.19-19.56)	16.64 (7.44~20.80)	11.87 (7.24~20.35)	12.84 (6.58~17.47)	22.55 (12.54~32.54)
Accuracy = 50% accuracy (95% CI)					
Gaussian[$]	48.28 (39.66~57.76)	7.14 (0.00~17.86)	35.71 (17.86~53.57)	66.67 (50.00~83.33)	66.67 (50.00~83.33)
SVM[$]	44.83 (36.21~53.45)	3.57 (0.00-10.71)	32.14 (14.29~50.00)	70.00 (53.33~86.67)	63.33 (46.67~80.00)

`

Random forest[$]	43.10 (34.48 ~52.59)	0	42.86 (25.00~60.71)	76.67 (60.00~90.00)	83.33 (66.67~96.67)
Simple[$]	51.72 (42.24~60.34)	0	28.57 (14.29~46.43)	80.00 (66.67~93.33)	80.00 (63.33~93.33)

* P=0.031, compared to random forest

** P<0.05, compared to normal, mild, and moderate groups with all models

$ For all models, p<0.05 according to OSA severity, by linear by linear association

[1] CI: confidence interval; [2] SVM: Support vector machine; [3] IQR: Interquartile range

`

Table 5. Clinical characteristics of patients.

Clinical characteristics	All (N = 423)	Training (N = 256)	Test (N = 167)	p-value
M:F	356:67	214:42	142:25	0.631
Age (years)	48.1 ± 12.8	47.9 ± 12.9	48.6 ± 12.7	0.509
BMI (kg/m^2)	27.0 ± 4.2	27.2 ± 4.2	26.7 ± 4.0	0.278
Time in bed (min)	479.9 ± 27.1	479.9 ± 27.5	479.9 ± 26.7	0.600
Sleep latency (min)	16.1 ± 22.9	16.7 ± 24.9	15.1 ± 19.7	0.925
Sleep efficiency (min)	79.9 ± 13.3	79.5 ± 14.0	80.4 ± 12.3	0.803
Apnea hypopnea index (events/h)	32.6 ± 24.4	32.9 ± 24.8	32.0 ± 23.8	0.816
Apnea index (events/h)	21.0 ± 21.8	21.8 ± 22.9	19.8 ± 20.2	0.568
Obstructive apnea index (/h)	18.3 ± 19.5	18.8 ± 20.3	17.4 ± 18.3	0.736
Normal/Mild/Moderate/Severe	43/80/109/191	26/46/72/110	17/34/37/81	0.505

BMI: body mass index, M: male, F: female

`

Table 6. Performance of binary classification for prediction of obstructive sleep apnea from sleep breathing sound based on three different apnea hypopnea index cut off values.

Threshold	Accuracy	Kappa	Sensitivity	Specificity	PPV	NPV	F1 score	PRC area	AUC
5 events/h	**88.166**	**0.459**	**90.79**	**64.71**	**95.83**	**44.00**	**0.932**	**0.989**	**0.909**
15 events/h	**82.249**	**0.579**	**87.29**	**70.59**	**87.29**	**70.59**	**0.873**	**0.950**	**0.890**
30 events/h	**81.657**	**0.632**	**82.95**	**80.25**	**82.02**	**81.25**	**0.807**	**0.907**	**0.896**

PPV: positive predictive value, NPV : negative predictive value, PRC area : area under precision recall curve, AUC : area under curve

`

Table 7. List of selected features in model 2. In model 2, noise reduced sound data and selected set of features were used as input data.

Origin features	**Description**[1]	**Selected features**
Fraction Of Low Energy Windows	This is a good measure of how much of a signal is quiet relative to the rest of a signal.	Fraction Of Low Energy Windows Overall Standard Deviation0
		Derivative of Fraction Of Low Energy Windows Overall Standard Deviation0
Linear Predictive Coding (LPC)	Spectral envelope based on the information of a linear predictive model	Derivative of LPC Overall Standard Deviation0
		Derivative of LPC Overall Standard Deviation1
		Derivative of LPC Overall Standard Deviation2
		Derivative of LPC Overall Standard Deviation3
		Derivative of LPC Overall Standard Deviation4
		Derivative of LPC Overall Standard Deviation5
		Derivative of LPC Overall Standard Deviation6
		LPC Overall Average4
		LPC Overall Average6
Mel Frequency Cepstrum Coefficients (MFCC)	A short-term power spectrum based on the nonlinear mel scale of frequency. This is concisely describes the overall shape of a spectral envelop, and is commonly used as feature in speech recognition and music information retrieval such as genre classification	Derivative of MFCC Overall Standard Deviation0
		Derivative of MFCC Overall Standard Deviation2
		Derivative of MFCC Overall Standard Deviation4
		Derivative of MFCC Overall Standard Deviation5
		Derivative of MFCC Overall Standard Deviation8
		MFCC Overall Average5
		MFCC Overall Average9
		Derivative of MFCC Overall Average3
		Derivative of MFCC Overall Average4
Spectral Rolloff Point	This is a measure of the amount of the right-skewedness of the power spectrum.	Spectral Rolloff Point Overall Standard Deviation0
Strength Of Strongest Beat	This is a measure of how strong the strongest beat is compared to other possible beats.	Strength Of Strongest Beat Overall Standard Deviation0
		Derivative of Strength Of Strongest Beat Overall Standard Deviation0
Strongest Frequency Via Fast Fourier Transform (FFT) Maximum	This feature provides the index of the maximum value in the power spectrum	Derivative of Strongest Frequency Via FFT Maximum Overall Average0
Strongest Frequency	This feature is an approximation of the pitch of the	Derivative of Strongest Frequency Via Zero Crossings Overall Average0

`

Via Zero Crossings	signal if the signal is monophonic	
Zero Crossings	This feature measures how many times the signal value crosses zero in a given window.	Derivative of Zero Crossings Overall Average0
Total number of features	**8**	**26**

`

Table 8. List of selected features in model 4.
In model 4, naïve recorded sound data without noise reduction and selected set of features were used as input data.

Origin features	**Description[1]**	**Selected features**
Area Method of Moments of ConstantQ-based Mel Frequency Cepstrum Coefficients (MFCC)	Area method of moments refers to numeric quantities at some distance from a refer- ence point or axis. ConstantQ-based MFCC implements an alternative to the MFCC that directly calculates the logarithmic frequency bins rather than performing a Fast Fourier Transform (FFT) and rebinning the content.	Area Method of Moments of ConstantQ-based MFCCs Overall Average0
Beat Histogram	This is histogram showing the strength of different rhythmic periodicities in a signal.	Beat Histogram Overall Standard Deviation35
		Beat Histogram Overall Standard Deviation36
		Beat Histogram Overall Standard Deviation71
		Beat Histogram Overall Standard Deviation72
Fraction Of Low Energy Windows	This is a good measure of how much of a signal is quiet relative to the rest of a signal.	Derivative of Fraction Of Low Energy Windows Overall Average0
		Derivative of Fraction Of Low Energy Windows Overall Standard Deviation0
		Fraction Of Low Energy Windows Overall Standard Deviation0
Linear Predictive Coding (LPC)	Spectral envelope based on the information of a linear predictive model	Derivative of LPC Overall Standard Deviation0
		Derivative of LPC Overall Standard Deviation3
		Derivative of LPC Overall Standard Deviation4
		Derivative of LPC Overall Standard Deviation5
Mel Frequency Cepstrum Coefficients (MFCC)	A short-term power spectrum based on the nonlinear mel scale of frequency. This is concisely describes the overall shape of a spectral envelop, and is commonly used as feature in speech recognition and music information retrieval such as genre classification	Derivative of MFCC Overall Average10
		Derivative of MFCC Overall Average12
		Derivative of MFCC Overall Average2
		Derivative of MFCC Overall Average3
		Derivative of MFCC Overall Average4
		Derivative of MFCC Overall Average9
Relative Difference Function	This feature calculates the log of the derivative of the root mean square. This is useful for onset detection.	Derivative of Relative Difference Function Overall Average0
		Derivative of Relative Difference Function Overall Standard Deviation0

`

Spectral Rolloff Point	This is a measure of the amount of the right-skewedness of the power spectrum.	Derivative of Spectral Rolloff Point Overall Average0
Strongest Frequency Via Fast Fourier Transform (FFT) Maximum	This feature provides the index of the maximum value in the power spectrum.	Derivative of Strongest Frequency Via FFT Maximum Overall Average0
Total number of features	**8**	**26**

`

Table 9. Performance of AHI estimation based on four regression models.

Models	Correlation coefficient	Mean absolute error	Root mean squared error
Model 1	**0.7744**	**11.5664**	**15.1921**
Model 2	**0.7747**	**11.2597**	**15.0192**
Model 3	**0.776**	**11.7586**	**15.1312**
Model 4	**0.7838**	**10.7924**	**14.7293**

Model 1: noise reduction without feature selection, Model 2: noise reduction with feature selection, Model 3: without noise reduction and feature selection, and Model 4: with feature selection without noise reduction.

`

Table 10. Multivariate analysis using linear regression model to identify factors associated with error (measured AHI – estimated AHI).

Input variables	B	95% CI
AHI	0.329	0.242, 0.415
Sleep efficiency	-0.197	-0.348, -0.046
BMI	0.490	-0.022, 1.001
Age	0.088	-0.064, 0.241
Sex (female)	1.148	-3.982, 6.278

Measured AHI, sleep efficiency, BMI, age and sex were the input variables.

AHI: apnea hypopnea index, BMI: body mass index, CI: confidence interval

`

References

1. Bradley TD and Floras JS, *Obstructive sleep apnoea and its cardiovascular consequences.* Lancet, 2009. **373**(9657): p. 82–93.
2. Engleman HM and Douglas NJ, *Sleep. 4: Sleepiness, cognitive function, and quality of life in obstructive sleep apnoea/hypopnoea syndrome.* Thorax, 2004. **59**(7): p. 618-22.
3. Fung JW, et al., *Severe obstructive sleep apnea is associated with left ventricular diastolic dysfunction.* Chest, 2002. **121**(2): p. 422-9.
4. Jin H, et al., *Acoustic Analysis of Snoring in the Diagnosis of Obstructive Sleep Apnea Syndrome: A Call for More Rigorous Studies.* J Clin Sleep Med, 2015. **11**(7): p. 765-71
5. Kim J, et al., *Exploiting temporal and nonstationary features in breathing sound analysis for multiple obstructive sleep apnea severity classification.* Biomed Eng Online, 2017. **16**(1): p. 6.
6. Kim JW, et al., *Prediction of Obstructive Sleep Apnea Based on Respiratory Sounds Recorded Between Sleep Onset and Sleep Offset.* Clin Exp Otorhinolaryngol, 2018. **12**(1): p. 72-78.
7. Kim T, Kim JW, and Lee K, *Detection of sleep disordered breathing severity using acoustic biomarker and machine learning techniques.* Biomed Eng Online, 2018. **17**(1): p. 16.
8. Frank E, et al., *Online appendix for "data mining: practical machine learning tools and techniques"*. 4 ed. 2016, Burlington: Morgan Kaufmann.
9. Liu X, et al., *Improving precision of glomerular filtration rate estimating model by ensemble learning.* J Transl Med, 2017. **15**(1): p. 231.

`

10. Stevens LA, Zhang Y, and Schmid CH, *Evaluating the performance of equations for estimating glomerular filtration rate.* J Nephrol, 2008. **21**(6): p. 797-807.
11. Efron B and Tibshirani RJ, *An Introduction to the Bootstrap*. 1993, New York, NY: Chapman & Hall.
12. Ben-Israel N, Tarasiuk A, and Zigel Y, *Obstructive apnea hypopnea index estimation by analysis of nocturnal snoring signals in adults.* Sleep, 2012. **35**(9): p. 1299-1305.
13. Fiz JA, et al., *Continuous analysis and monitoring of snores and their relationship to the apnea-hypopnea index.* Laryngoscope, 2010. **120**(4): p. 854-62.
14. Nakano H, et al., *Monitoring sound to quantify snoring and sleep apnea severity using a smartphone: proof of concept.* Monitoring sound to quantify snoring and sleep apnea severity using a smartphone: proof of concept, 2014 **10**(1): p. 73-78
15. Xu H, et al., *Nocturnal snoring sound analysis in the diagnosis of obstructive sleep apnea in the Chinese Han population.* Sleep Breath, 2015. **19**(2): p. 599-605.
16. Polikar R, *Ensemble based systems in decision making.* IEEE Circuits Syst Mag, 2006. **6**(3): p. 21-45.
17. Ayappa I, et al., *Immediate consequences of respiratory events in sleep disordered breathing.* Sleep Med, 2005. **6**(2): p. 123-30.
18. Fujinaga I, *Adaptive optical music recognition*. 1997, McGill University.
19. Elwali A and Moussavi Z, *Determining Breathing Sound Features Representative of Obstructive Sleep Apnea During Wakefulness with Least Sensitivity to Other Risk Factors.* J Med Biol Eng, 2018(In

press).

20. Kim JW, et al., *Relationship Between Snoring Intensity and Severity of Obstructive Sleep Apnea.* Clin Exp Otorhinolaryngol, 2015. **8**(4): p. 376-80.
21. Beaudin, A.E., et al., *Impact of obstructive sleep apnoea and intermittent hypoxia on cardiovascular and cerebrovascular regulation.* Exp Physiol, 2017. **102**(7): p. 743-763.
22. Fietze, I., et al., *Long-term variability of the apnea-hypopnea index in a patient with mild to moderate obstructive sleep apnea.* J Clin Sleep Med, 2020. **16**(2): p. 319-323.
23. Sforza, E., et al., *Internight Variability of Apnea-Hypopnea Index in Obstructive Sleep Apnea Using Ambulatory Polysomnography.* Front Physiol, 2019. **10**: p. 849.
24. McCall, C. and W.V. McCall, *Objective vs. subjective measurements of sleep in depressed insomniacs: first night effect or reverse first night effect?* J Clin Sleep Med, 2012. **8**(1): p. 59-65.
25. Kundel, V. and N. Shah, *Impact of Portable Sleep Testing.* Sleep Med Clin, 2017. **12**(1): p. 137-147.
26. Jin, H., et al., *Acoustic Analysis of Snoring in the Diagnosis of Obstructive Sleep Apnea Syndrome: A Call for More Rigorous Studies.* J Clin Sleep Med, 2015. **11**(7): p. 765-71.
27. Nakano, H., et al., *Monitoring sound to quantify snoring and sleep apnea severity using a smartphone: proof of concept.* J Clin Sleep Med, 2014. **10**(1): p. 73-8.
28. Kim, J.W., et al., *Prediction of Obstructive Sleep Apnea Based on Respiratory Sounds Recorded Between Sleep Onset and Sleep Offset.* Clin Exp Otorhinolaryngol, 2019. **12**(1): p. 72-78.

`

29. Kim, J.W., et al., *Prediction of Apnea-Hypopnea Index Using Sound Data Collected by a Noncontact Device.* Otolaryngol Head Neck Surg, 2020. **162**(3): p. 392-399.
30. *Sleep-related breathing disorders in adults: recommendations for syndrome definition and measurement techniques in clinical research. The Report of an American Academy of Sleep Medicine Task Force.* Sleep, 1999. **22**(5): p. 667-89.
31. Team, A. *Audacity®: Free Audio Editor and Recorder [Computer application]*. Available from: https://audacityteam.org.
32. McEnnis D, et al., *jAudio: An Feature Extraction Library.* Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), 2005: p. 600-603.
33. Frank E, Hall MA, and Witten IH, *The WEKA Workbench*. 2016, Morgan Kaufmann: Burlington, MA.
34. Chawla NV, et al., *SMOTE: Synthetic Minority Over-sampling Technique.* Journal of Artificial Intelligence Research, 2002. **16**: p. 321-257.
35. Hall MA, *Correlation-based feature subset selection for machine learning*. 1998, University of Waikato: Hamilton.
36. Lundberg, S.M., G.G. Erion, and S.-I. Lee, *Consistent Individualized Feature Attribution for Tree Ensembles.* ArXiv, 2018. **abs/1802.03888**.
37. Park, S.H. and H.Y. Kressel, *Connecting Technological Innovation in Artificial Intelligence to Real-world Medical Practice through Rigorous Clinical Validation: What Peer-reviewed Medical Journals Could Do.* J Korean Med Sci, 2018. **33**(22): p. e152.
38. Jonas, D.E., et al., *Screening for Obstructive Sleep Apnea in Adults:*

`

Evidence Report and Systematic Review for the US Preventive Services Task Force. Jama, 2017. **317**(4): p. 415-433.

39. Benjafield, A.V., et al., *Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis.* Lancet Respir Med, 2019. **7**(8): p. 687-698.
40. Bland, M., *Assessing agreement using Cohen's kappa*. fourth ed. An introduction to medical statistics 2015: Oxford University Press.
41. Abrahamyan, L., et al., *Diagnostic accuracy of level IV portable sleep monitors versus polysomnography for obstructive sleep apnea: a systematic review and meta-analysis.* Sleep Breath, 2018. **22**(3): p. 593-611.
42. Papini, G.B., et al., *Wearable monitoring of sleep-disordered breathing: estimation of the apnea-hypopnea index using wrist-worn reflective photoplethysmography.* Sci Rep, 2020. **10**(1): p. 13512.
43. Kapur, V.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**(3): p. 479-504.
44. Isabelle Guyon and Andre Elisseeff, *An Introduction to Variable and Feature Selection.* Journal of Machine Learning Research 3, 2003: p. 1157-1182
45. Thorn Thomas, *Background Noise Reduction: One of Your Smartphone's GreatestTools*, in *TechRadar*. 2014, Future US Inc: Bath, England, UK.
46. M. Jeub, et al., *Noise reduction for dual-microphone mobile phones exploition power leveL difference.* 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),

2012: p. 1693-1696.

47. Kim, T., J.W. Kim, and K. Lee, *Detection of sleep disordered breathing severity using acoustic biomarker and machine learning techniques.* Biomed Eng Online, 2018. **17**(1): p. 16.

48. Bountourakis, V., L. Vrysis, and G. Papanikolaou, *Machine Learning Algorithms for Environmental Sound Recognition: Towards Soundscape Semantics,* in *Proceedings of the Audio Mostly 2015 on Interaction With Sound*. 2015, Association for Computing Machinery: Thessaloniki, Greece. p. Article 5.

49. Dafna, E., A. Tarasiuk, and Y. Zigel, *Sleep staging using nocturnal sound analysis.* Sci Rep, 2018. **8**(1): p. 13474.

50. Levartovsky, A., et al., *Breathing and Snoring Sound Characteristics during Sleep in Adults.* J Clin Sleep Med, 2016. **12**(3): p. 375-84.

51. Akhter, S., et al., *Snore Sound Analysis Can Detect the Presence of Obstructive Sleep Apnea Specific to NREM or REM Sleep.* J Clin Sleep Med, 2018. **14**(6): p. 991-1003.

52. Koh, T.K., et al., *Snoring Sound Intensity and Formant Frequencies by Sleep Position in Patients with Positional Obstructive Sleep Apnea.* Korean J Otorhinolaryngol-Head Neck Surg, 2020. **63**(7): p. 308-313.

53. Zaffaroni, A., et al., *Sleep Staging Monitoring Based on Sonar Smartphone Technology.* Annu Int Conf IEEE Eng Med Biol Soc, 2019. **2019**: p. 2230-2233.

54. Kim, D.K., et al., *Rethinking AASM guideline for split-night polysomnography in Asian patients with obstructive sleep apnea.* Sleep Breath, 2015. **19**(4): p. 1273-7.

`

국문 초록

서론: 수면 중 호흡음은 폐쇄성 수면 무호흡증(Obstructive Sleep Apnea, OSA)의 잠재적인 바이오마커로 간주되어 왔다. 수면 중 숨소리는 대부분의 스마트폰 장치에 있는 마이크를 사용하여 쉽게 녹음할 수 있기 때문에 OSA 의 사전 스크리닝 목적의 평가 도구로 쉽게 구현될 수 있다. 본 연구는 스마트폰에 녹음된 소리를 이용하여 OSA를 예측하고 소음처리 및 선택된 소리의 feature 에 대한 최적의 설정을 파악하고자 한다.

방법: 2015년 8월부터 2019년 8월까지 코골이 또는 수면무호흡증으로 상급종합병원 수면센터를 방문한 환자를 대상으로 단면연구를 수행 하였다. 수면 중 오디오 녹음은 일상적인 밤새 실험실 내 수면다원검사 중에 스마트폰을 사용하여 수행되었다. 총 423명의 환자를 분석하였고, 데이터는 train set (60%, n = 256)과 test set(40%, n = 167)으로 분할되었다. 랜덤 포레스트 알고리즘을 사용하여 무호흡 저호흡 지수(apnea hypopnea index, AHI)의 임계값 5, 15 또는 30 회/시간에 따라 세 가지 기준에 대해 이진 분류를 별도로 수행하였고, 더불어 실제 AHI를 예측하기 위해 입력 사운드에서 노이즈 감소 및 기능 선택에 따라 다음과 같이 4개의 회귀 모델을 생성하였다. 1) 특징 선택 없이 잡음 감소, 2) 특징 선택으로 잡음 감소, 3) 잡음 감소 및 특징 선택 없이, 4) 잡음 감소 없이 특징 선택. 또한 예측오류에 영향을 미칠 수 있는 임상 및 수면다원검사상의 여러가지 변수들을 평가하였다.

결과: AHI 임계값 5, 15 및 30개 회/시간에 대한 예측 정확도는 88.16%, 82.25% 및 81.66%였으며 곡선 아래 영역은 각각 0.9, 0.89 및 0.9였다. 회귀분석에서는 잡음이 제거되지 않고 선택된 feature만 있는 녹음된 소리를 사용할 경우 상관계수가 가장 높았다(r=0.784, 95% 신뢰구간(CI): 0.689~0.879). 환자의 실제 AHI(베타=0.329, 95% CI: 0.242~0.415)와 수면효율(베타=−0.197, 95% CI: −0.348~−0.046)은 추정오차와 관련이 있는 것으로 나타났다.

결론: 스마트폰을 이용하여 녹음한 수면 호흡음을 통해 OSA을

`

예측할 수 있다. 향후 연구는 다양한 스마트폰 기기를 사용한 실생활 녹음에 초점을 맞춰야 한다

주요어 ：Obstructive sleep apnea; sleep breathing sound; smartphone; prediction

학　　번：**2018-39428**