Original Research

DIGITAL HEALTH

DIGITAL HEALTH
Volume 9: 1–19

DOI: 10.1177/20552076231211550
journals.sagepub.com/home/dhj

Sage

Novel method combining multiscale attention entropy of overnight blood oxygen level and machine learning for easy sleep apnea screening

Zilu Liang

Abstract

Objective: Sleep apnea is a common sleep disorder affecting a significant portion of the population, but many apnea patients remain undiagnosed because existing clinical tests are invasive and expensive. This study aimed to develop a method for easy sleep apnea screening.

Methods: Three supervised machine learning algorithms, including logistic regression, support vector machine, and light gradient boosting machine, were applied to develop apnea screening models at two apnea–hypopnea index cutoff thresholds: ≥ 5 and ≥ 30 events/hours. The SpO2 recordings of the Sleep Heart Health Study database ($N = 5786$) were used for model training, validation, and test. Multiscale entropy analysis was performed to derive a set of multiscale attention entropy features from the SpO2 recordings. Demographic features including age, sex, body mass index, and blood pressure were also used. The dependency among the multiscale attention entropy features were handled with the independent component analysis.

Results: For cutoff ≥ 5/hours, logistic regression model achieved the highest Matthew's correlation coefficient (0.402) and area under the curve (0.747), and reasonably good sensitivity (75.38%), specificity (74.02%), and positive predictive value (92.94%). For cutoff ≥ 30/hours, support vector machine model achieved the highest Matthew's correlation coefficient (0.545) and area under the curve (0.823), and good sensitivity (82.00%), specificity (82.69%), and negative predictive value (95.53%).

Conclusions: Our models achieved better performance than existing methods and have the potential to be integrated with home-use pulse oximeters.

Keywords

Sleep apnea, SpO2, oximeter, machine learning, multiscale entropy, attention entropy, imbalanced classification

Submission date: 30 May 2023; Acceptance date: 16 October 2023

Introduction

Sleep apnea is a common sleep disorder characterized by frequent episodes of partial or full blockage of the respiratory tract during sleep. Sleep apnea is clinically diagnosed using the apnea–hypopnea index (AHI), which is defined as the number of apnea or hypopnea events per hour during a night's sleep. A person is considered as having sleep apnea if the AHI is larger than or equal to 5. Continuous positive airway pressure (CPAP) treatment is often recommended for people with an AHI larger than 30 in Japan.[1] The symptoms of sleep apnea include loud

Kyoto University of Advanced Science (KUAS), Japan

Corresponding author:
Zilu Liang, Ubiquitous and Personal Computing Lab, Faculty of Engineering, Kyoto University of Advanced Science (KUAS), 18 Yamanouchi Gotanda-cho, Ukyo-ku, Kyoto 615-8577, Japan.
Email: liang.zilu@kuas.ac.jp

snoring and excessive daytime sleepiness. In the long-term, sleep apnea is associated with increased risk of diabetes, hypertension, depression, and memory loss.[2,3] It is, therefore, important for sleep apnea patients to seek timely diagnosis and treatment.

Clinical diagnosis of sleep apnea requires a patient to undergo an overnight polysomnography (PSG) test in a hospital or sleep clinic. A PSG test monitors many physiological signals during sleep to gain a holistic view of a patient's physical states during sleep. The set of signals is then manually scored by a registered sleep technician. PSG test is expensive and time-consuming, which limits its accessibility and affordability as a method for sleep apnea screening. To address the problem, many studies have attempted to develop more affordable sleep apnea screening systems. These systems aim to achieve reasonable accuracy in detecting sleep apnea episodes using only a subset of the signals measured in a PSG test such as electroencephalogram (EEG)[4] and electrocardiogram (ECG).[5] In a related vein, some studies developed sleep apnea screening methods that rely on the information stored in electronic health records (EHRs) such as demographic characteristics, laboratory blood test results, comorbidities, medication, questionnaire scores, etc.[6–10,12,11,13] Although EHR-based methods are promising, they are not applicable to at-home sleep apnea screens because not all the required data are readily available to individuals.

As consumer sleep tracking technologies burgeon, research interest has started to shift towards developing new sleep analysis models that are compatible with the sensing modalities available in those devices.[14–16] Most of the consumer sleep trackers nowadays have embedded photoplethysmography sensors that continuously measure users' peripheral oxygen saturation (SpO2), making it an ideal sensing modality for sleep apnea detection in free-living environments. Many studies in this line of research formulated sleep apnea detection into an epoch-wise classification problem.[17–21] A whole night's SpO2 recording was segmented into small intervals called epochs. The size of an epoch was often set to 1 minute. The constructed classification model was supposed to map each epoch to either a positive or negative apnea event. This scheme for apnea event classification suffers two major drawbacks. First, the signal segmentation process introduces dependency among training samples, violating the assumption of independent and identical distribution that is central to many machine learning techniques. Despite the seemingly large number of segments, the degree of freedom is limited by the cohort size, which is often too small in prior studies.[26] For example, the widely used Apnea-ECG dataset only contains eight recordings of SpO2 signals, while the St. Vincent's University Hospital/University College Dublin Sleep Apnea Database contains SpO2 recordings of only 25 subjects. The models thus may suffer from overfitting and may not generalize well to new data. Second, the performance evaluation of the existing model often centers on the detection of individual apnea events rather than an overall evaluation of whether a user is likely to have sleep apnea. While one may argue that the accurate detection of respiratory events allows for the calculation of AHI (and thus sleep apnea screening), the evaluation of the latter is mostly missing in prior studies and thus leaving it unknown in terms of the models' performance in apnea screening. A few recent studies developed SpO2-based apnea screening models and validated them on larger datasets. Trained with 18 features and validated on the São Paulo Epidemiologic Sleep Study dataset (N = 887), the OxyDOSA model achieved an area under the receiver operating characteristic curve (AUC) of 0.94 and an accuracy of 86%.[22] In another study, a least-squares boosting (LSBoost) model was trained on 32 features and validated on the Sleep Heart Health Study (SHHS) dataset as well as the Rio Hortega University Hospital (RHUH) dataset.[23] The LSBoost model achieved an accuracy between 87.23% and 96.58% for a binary classification. In a recent study, a convolutional neural network-based deep learning model was developed by Sharma et al.[24] to achieve segment-wise classification of sleep apnea events. This model was validated on the SHHS dataset and achieved an AUC of 90.4 and an accuracy 82.2%. Another deep learning model OxiNet was recently developed by Levy et al.[25] to predict subject-wise AHI values. This model was validated on six large datasets (N = 369–5778) and yielded an average $F1$-score between 0.75 and 0.84.

In this article, we developed a new method for sleep apnea screening based on analyzing overnight SpO2 recordings. Our method combines multiscale entropy (MSE) analysis and machine learning. The problem of interest was formulated into a binary classification problem. We were interested in distinguishing the positive class and the negative class at two AHI cutoffs: ≥ 5 and ≥ 30/hours. The first cutoff is the clinical diagnostic criteria of sleep apnea, and it allows the detection of sleep apnea patients from healthy population. The second cutoff allows the identification of people with severe apnea, and it is often used to decide whether a patient needs CPAP treatment in Japan.[1]

The key innovation of our method is the multiscale attention entropy (MSAE)—a novel way to construct features that characterize the complexity of the SpO2 signals for various temporal scales. MSE analysis generates insights into the temporal fluctuation of the information encoded in the SpO2 signals. We used a new entropy measure—attention entropy—as the base entropy in the MSE analysis. Attention entropy has advantages over other conventional entropy measures such as sample entropy and approximate entropy because it is robust to the length of the signal and is parameter-free. The attention entropy of the SpO2 signals computed for various scale

factors formed the raw feature set, which was scaled and then processed using the independent component analysis (ICA) to generate independent features for model construction. We considered both short, medium, and long-term time scales of up to 30 minutes in the MSAE analysis.

The contribution of this article is as follows. First, we proposed a new method that combines the MSAE and ICA to derive independent features that characterize the complexity of the overnight SpO2 signals at short-, medium-, and long-term temporal scales. Second, we examined the discriminating power of the attention entropy for a wide range of time scale from 1 second to 30 minutes. To the best of our knowledge, this is the first study that investigates the multiscale entropy for time scales longer than 1 minute. Our analysis found that the attention entropy of the overnight SpO2 signals at time scale longer than 1 minute could be useful features for distinguishing apnea positive and negative. This is a new finding that no prior study had discovered. Third, we developed and validated classification models for sleep apnea screening that will have greater compatibility with home use sleep tracking technologies, because the models only rely on one input signal—overnight SpO2—that is nowadays readily measurable with off-the-shelf digital sleep health gadgets.

The rest of the article is organized as follows. The "Method" section explains the database, feature construction and transformation, model training and test, and performance evaluation measures. The "Result" section shows the visualization of the MSAE for the positive and negative groups for various scale factors and the performance of the apnea screening models. In the "Discussion" section, we provide interpretations of the principal findings, compare our work with prior work, and highlight the limitations of the current work. The article is concluded in the last section.

Method

Database

In this study, we used the SHHS database. The SHHS is a multi-center cohort study that aims to investigate the associations between sleep apnea and other diseases including stroke, hypertension, coronary heart disease, and all-cause mortality. Subjects who met the inclusion criteria were those 40 years or older with no history of treatment of sleep apnea and tracheostomy, and were not under home oxygen therapy at the time of the experiment. Subjects underwent two PSG tests with the second test conducted at least 3 years after the first one. PSG tests were performed at subjects' homes for better ecological validity. The SpO2 signals were recorded using a Nonin XPOD 3011 with an 8000J sensor attached to a finger. Recordings shorter than 4 hours were removed. The sampling rate was 1 Hz. All subjects provided written informed consent before experiments started. Access to the SHHS database was granted by the National Sleep Research Resource (NSRR), and the handling of the data in this study was compliant with the Data Access and Use Agreement (DAUA).

We used 5786 SpO2 recordings of subjects' first visits to build the sleep apnea screening models in this study. The recordings of the second visits were discarded to avoid introducing within-subject dependency. The demographics and sleep-related characteristics of the subjects are shown in Table 1. The mean age was 63.1 years and 47.7% were men. The apnea prevalence was high (82.6%) due to the intentional oversampling of snorers in SHHS. The prevalence of severe apnea as defined by AHI≥ 30/hours was 17.4% in the whole cohort and 20.6% among apnea patients. Compared to subjects without apnea, those with apnea as defined by cutoff ≥ 5/hours were older, more obese, had a higher percentage of men, had higher systolic blood pressure, slept shorter, had longer wakefulness after sleep onset, and had lower sleep efficiency. Similar tendency was observed for cutoff ≥ 30/hours. The positive and negative classes are imbalanced, with a ratio of 4.77:1 (for AHI ≥ 5/hours) and 1:4.87 (for AHI ≥ 30/hours). Further details of the SHHS can be obtained by Quan et al.[27]

Data were retrieved from the NSRR repositories upon approval.[28] The PSG recordings were downloaded as European Data Format (EDF) files from which the SpO2 signals were extracted to derive MSAE features. In line with prior studies,[17,29,30,23] we considered the followings as artifacts and removed them from the SpO2 signals: (1) readings of zeros, (2) readings below 50% or above 100%, and (3) sudden changes of more than 4% between consecutive readings. In addition to extracting the SpO2 signals from the PSG recordings in the SHHS database, we also used several demographic variables that individuals can easily obtain at home. Those variables include age, sex, body mass index (BMI), systolic blood pressure (BPS), and diastolic blood pressure (BPD).

It is worth noting that the database contains several harmonized AHI variables. This is because several major revisions were made to the guideline of sleep scoring rules along the years, leading to slight changes in the calculation of AHI. Based on the latest version of the guideline,[31] some of the AHI variables are considered as "recommended," while others are considered as "alternative" or "acceptable." We used the *nsrr_ahi_ph3r_aasm15* variable as the ground truth because the annotation rules are consistent with the recommended rules in the latest guideline.[31] In detail, the AHI was calculated as the number of apnea and hypopnea events with more than 30% nasal airflow reduction and more than 3% oxygen desaturation with or without arousal per hour of sleep. The calculation of AHI took into account both obstructive sleep apnea and central sleep apnea events. For apnea and non-apnea classification, *nsrr_ahi_ph3r_aasm15* ≥ 5 was mapped to the positive

Table 1. Demographics and sleep characteristics of subjects.

	All	Positive (≥ 5/hours)	Negative (< 5/hours)	Positive (≥ 30/hours)	Negative (< 30/hours)
No. of subjects	5786	4784	1002	986	4800
No. of males (%)	2758 (47.7)	2512 (52.5)	246 (24.6)	682 (69.2)	2076 (43.3)
Age[a]	63.1±11.2	64.0±11.0	58.7±11.3	65.7±10.6	62.6±11.3
BMI[a] (kg/m^2)	28.2±5.1	28.6±5.1	25.9±4.2	30.7±5.6	27.6±4.8
BPS[b]	127.4±19.3	128.2±19.2	123.4±19.5	130.9±18.9	126.6±19.3
BPD[c]	73.7±11.6	73.8±11.8	73.0±10.6	75.3±12.7	73.4±11.4
TST[d] (minutes)	359.9±64.5	356.9±64.8	374.2±61.0	345.3±67.2	362.9±63.5
WASO[e] (minutes)	61.5±44.0	64.0±45.1	49.5±36.0	74.7±50.4	58.7±42.1
SE[f] (%)	82.8±10.5	82.2±10.8	85.6±8.9	79.8±11.9	83.4±10.1
AHI[g] (events/hours)	17.9±16.1	21.1±16.0	2.9±1.3	46.6±15.8	12.1±7.6

[a]Body mass index.
[b]Systolic blood pressure.
[c]Diastolic blood pressure.
[d]Total sleep time.
[e]Wake after sleep onset.
[f]Sleep efficiency.
[g]Apnea-hypopnea index.

class corresponding to apnea, and *nsrr_ahi_ph3r_aasm15* < 5 was mapped to the negative class corresponding to non-apnea. For the classification of severe apnea and all other cases, *nsrr_ahi_ph3r_aasm15* ≥ 30 was mapped to the positive class corresponding to severe apnea, and *nsrr_ahi_ph3r_aasm15* < 30 was mapped to the negative class corresponding to all other cases. Figure 1 shows the examples of the SpO2 waveform for subjects with no apnea, mild to moderate apnea, and severe apnea, respectively.

Feature construction based on multiscale attention entropy analysis

In our proposed method, features were derived based on the MSE analysis of the overnight SpO2 recordings. The MSE analysis is a technique to evaluate the complexity and regularity of a signal at multiple time scales.[32] It generates insights into the dynamic temporal fluctuations of the information encoded in a signal and provides additional useful information that conventional single-value entropy measure fails to capture. The MSE analysis iterates over two steps for each specified scale factor τ: signal graining and entropy calculation. The originality of our method manifests in both steps. We applied attention entropy in place of the conventional sample entropy or approximate entropy to eliminate the need for phase space reconstruction and expensive parameter tuning. Furthermore, we considered large-scale factors representing much longer temporal scales than those used in prior studies. The process for calculating MSAE is illustrated in Figure 2. In what follows, we explain the MSAE analysis and feature construction in detail.

The first step of the MSAE analysis is to segment a cleaned SpO2 signal into non-overlapping coarse-grained sequences for different temporal scales. Given a digital SpO2 signal $x(i) = \{x(1),\ x(2),\ \ldots, x(N)\}$ $(i = 1, 2, \ldots, N)$, the coarse-grained signal for scale factor τ $(\tau \in \mathbb{N}+)$, denoted as $x_g^\tau(j) = \{x_g^\tau(1),\ x_g^\tau(2),\ \ldots, x_g^\tau(N/\tau)\}$ $(j = 1, 2, \ldots, N/\tau)$, can be calculated by averaging all the data points within the j-th graining window, as shown in equation (1). For $\tau = 1$, $x_g^\tau(j)$ is equivalent to the original signal. For $\tau > 1$, the length of the grained signal reduces progressively as the scale factor τ increases. The upper bound of the scale factor was often heuristically set to a value between 10 and 50.[33–35] Given the high sampling rate of signals, the scale factors applied in previous studies correspond to very short temporal scales that are often less than 1 minute. In the present study, we set the maximum scale factor to 1800, which corresponds to a temporal scale of

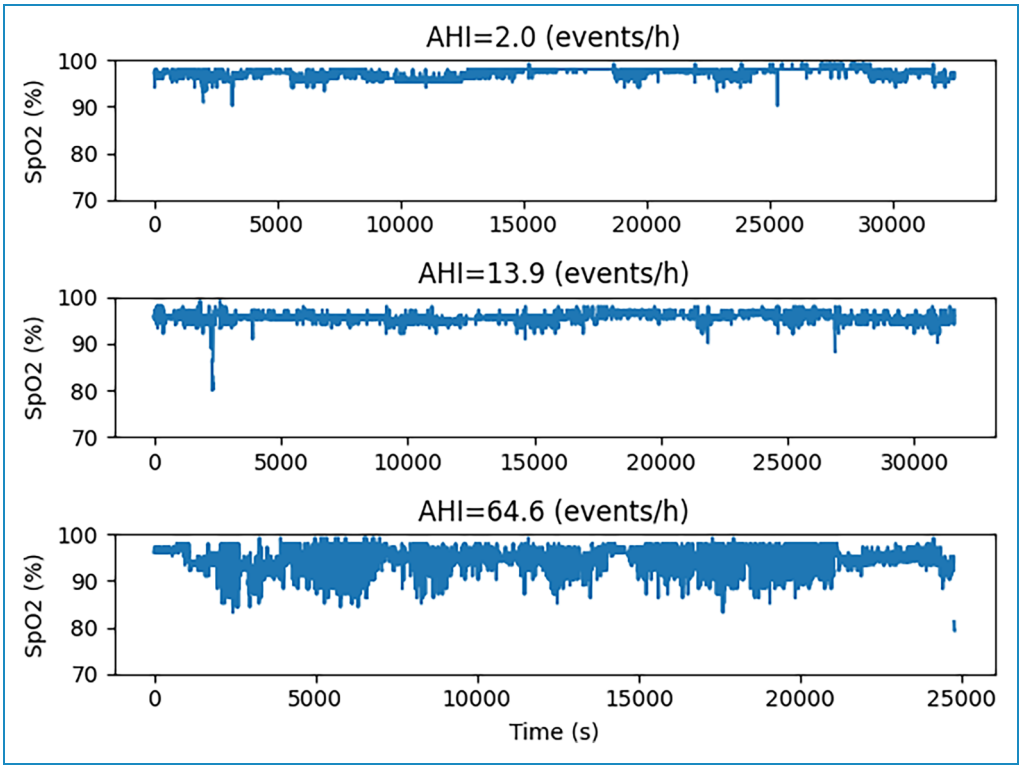

Figure 1. Examples of SpO2 signals of (upper) a subject without sleep apnea, (middle) a subject with mild to moderate apnea, and (lower) a subject with severe apnea.

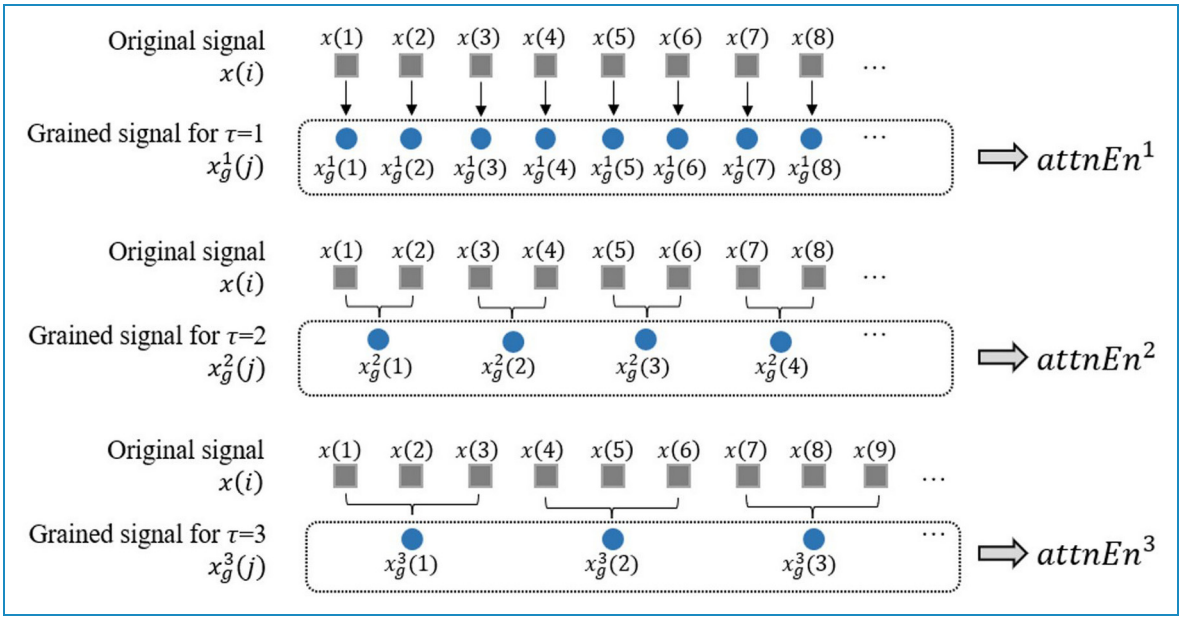

Figure 2. Process for multiscale attention entropy (MSAE) calculation.

30 minutes, to extract useful information about the signal complexity at short, medium, and long time scales.

$$x_g^{\tau}(j) = \frac{1}{\tau} \sum_{i=(j-1)\tau+1}^{j\tau} x(i), \quad 1 \le j \le \frac{N}{\tau} \tag{1}$$

The second step of the MSAE analysis is to calculate the entropy of the coarse-grained signal $x_g^{\tau}(j)$ for each scale factor τ ($\tau \in \mathbb{N}+$). Different types of entropy measures can be used in this step. Sample entropy and approximate entropy are the most used so far. However, those entropy measures have several limitations including lacking robustness to short time series and requiring intensive parameter tuning. In this study, we used a new parameter-free entropy measure called attention entropy (denoted as *attnEn*). The attention entropy characterizes the frequency distribution of the intervals between successive peak points in a time series.[36] It is robust to both short and long time series as it does not focus on the frequency distribution of all observations, and it saves the trouble of parameter tuning that is often required in the computation of conventional entropy measures. The calculation of attention entropy for a coarse-grained SpO2 signal $x_g^{\tau}(j)$ takes three steps. First, the peak points in $x_g^{\tau}(j)$ will be detected, which represent local maxima and local minima. Second, the intervals between two successive peak points for each pattern ω in $\mathbf{\Omega}$, denoted as $I_{\omega}^{\tau}(k)$, will be calculated. There are four key patterns in $\mathbf{\Omega}$: local maxima to local maxima, local minima to local maxima, local maxima to local minima, and local minima to local minima. The Shannon entropy

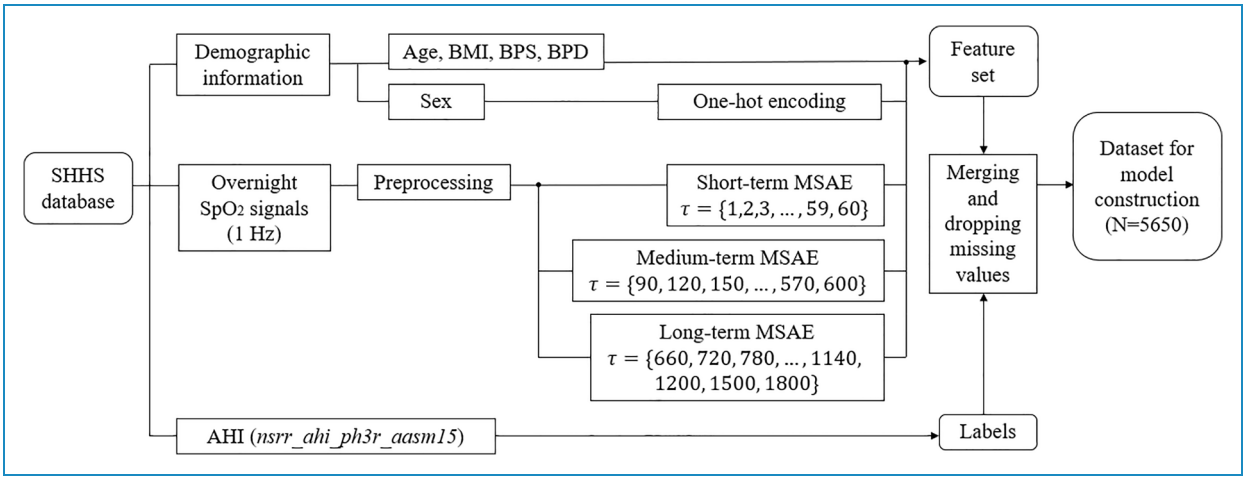

Figure 3. Process for feature and label construction. BMI: body mass index; BPS: systolic blood pressure; BPD: diastolic blood pressure.

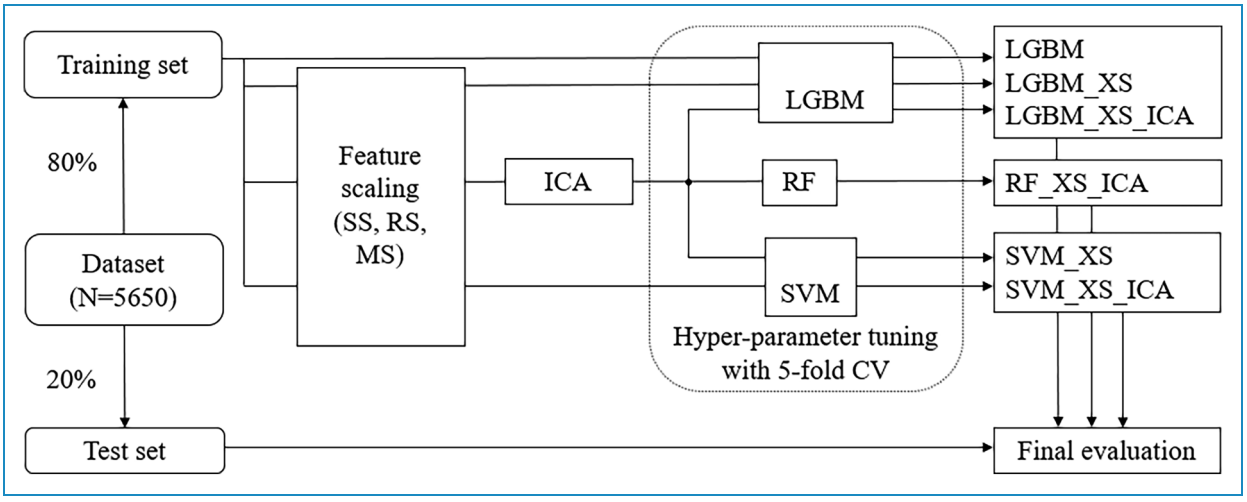

Figure 4. Process for model construction. SS: standard scaling; RS: robust scaling; MS: min-max scaling; XS refers to either SS, RS, or MS.

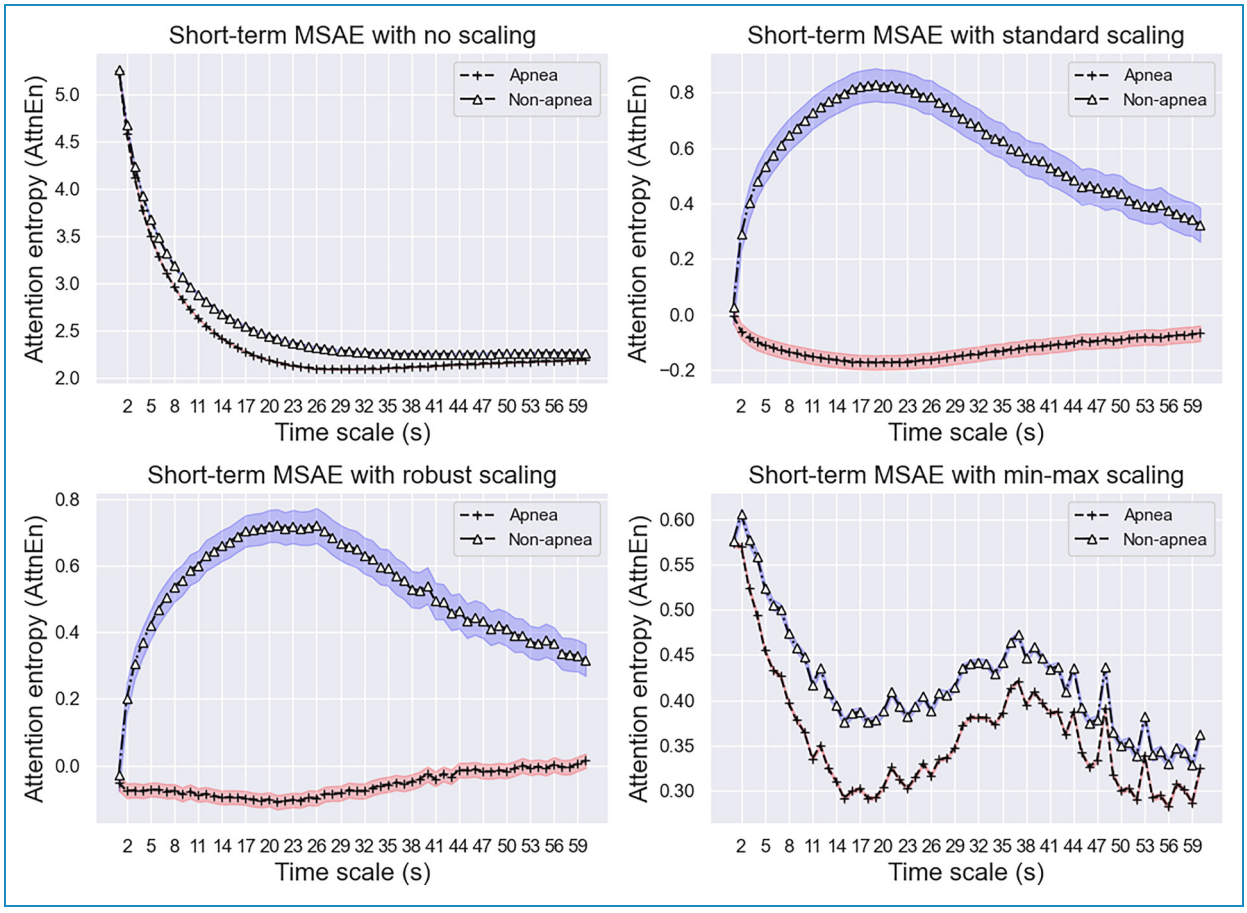

Figure 5. Multiscale attention entropy (MSAE) for short-term scale factor (cutoff $\geq$ 5/hours).

of the intervals for each pattern will be calculated separately, and then averaged to generate the attention entropy of the time series for scale factor τ, as shown in equation (2). Eventually, the feature set $S = \{attEn^{\tau}\}$ consists of the attention entropy of an SpO2 recording for different time scales. The values of τ for short-time scale were between 1 and 60 with an increment of 1, corresponding to a time range of 1–60 seconds. The values for medium-time scale were between 90 and 600 with an increment of 30, corresponding to a time range of 1.5–10 minutes. The values for long-time scale were between 660 and 1800, with an increment of 60 between 660 and 1200 and an increment of 300 between 1200 and 1800. The long-term scales correspond to 11–30 minutes.

$$attnEn^{\tau} = -\frac{1}{4}\sum_{\omega\in\Omega}\sum_{k} P(I_{\omega}^{\tau}(k))\log P(I_{\omega}^{\tau}(k)) \qquad (2)$$

In addition to the MSAE features, we also used demographic features including age, sex, BMI, BPS, and BPD. These features were shown to be associated to sleep apnea[38,39] and are easily attainable. Figure 3 illustrates the overall process for feature and label construction.

Feature transformation using independent component analysis

A potential issue with the MSAE-based feature construction is that the derived features are correlated. As feature dependency may compromise the performance of classification models, we applied the ICA to transform the raw features into a new set of features so that the statistical dependency among the transformed features is minimized. The ICA technique was originally used in the field of computational neural science for de-noising EEG signals[40] and was later applied to a variety of problems.[41] In this study, the input to the ICA is the original set of the MSAE features, and the output of the ICA is a set of maximally independent features which are linear combinations of the original MSAE features. The number of components was set to the dimension of the original feature set, as the objective of the ICA was to remove the dependencies among the original features instead of dimension reduction.

We firstly normalized the original MSAE features because the ICA is sensitive to the scale of the input.

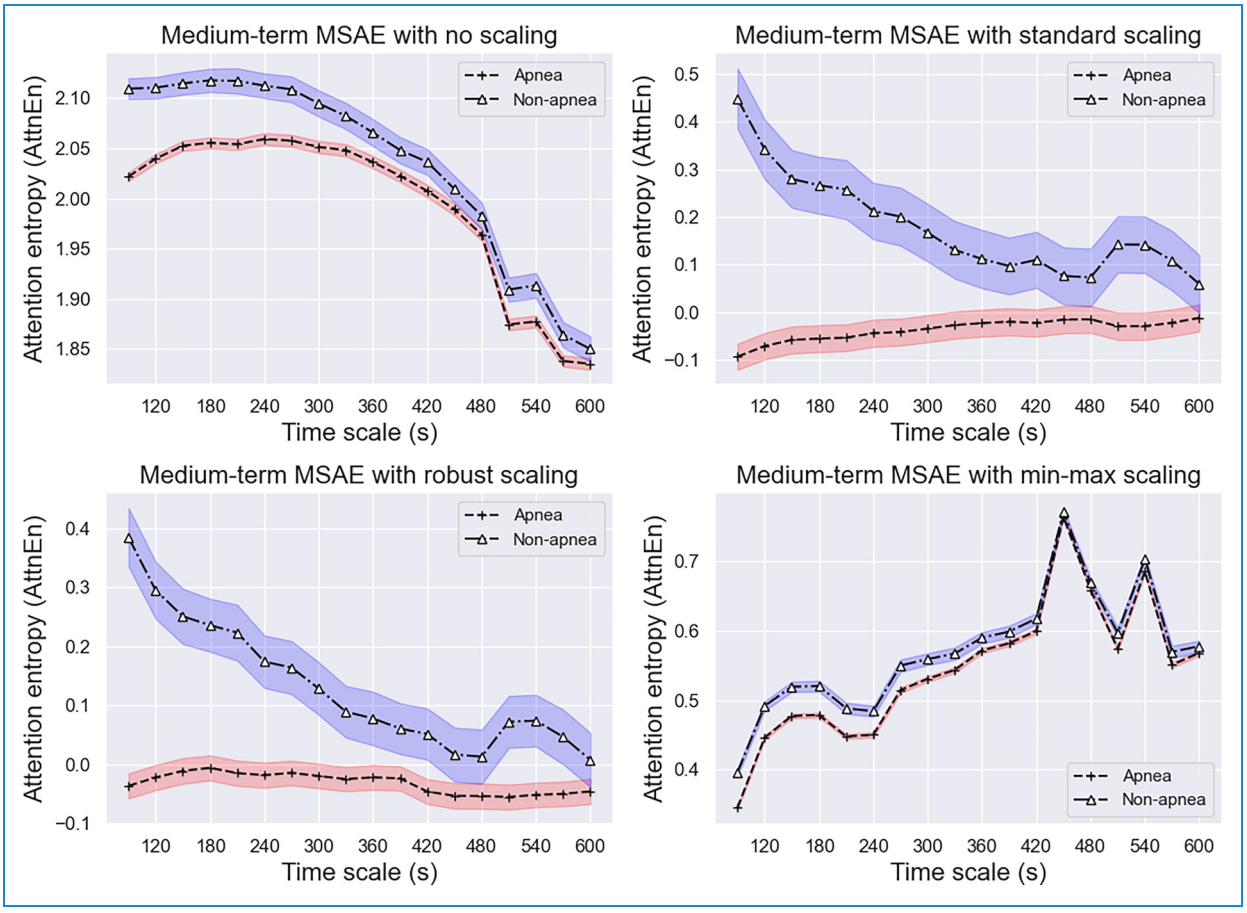

Figure 6. Multiscale attention entropy (MSAE) for medium-term scale factor (cutoff ≥ 5/hours).

Feature scaling is also an important preprocessing step for many machine learning algorithms, especially for distance-based learning algorithms such as support vector machine (SVM).[42] We normalized the range of the features using one of the following three methods.

(i) Standard scaling (SS) normalizes features to have zero mean and unit variance.
(ii) Robust scaling (RS) removes the median and scales features to the range between the first and the third quartiles. This scaling method is robust to outliers.
(iii) Min–max scaling (MS) normalizes features to between zero and one.

Model training, validation, and test

We applied three supervised machine learning algorithms that are most suited for binary classification with medium data size: logistic regression (LR), SVM, and light gradient boosting machine (LGBM). LR and SVM have previously shown promise in sleep apnea classification with demographic features,[9,43] whereas no prior study has used LGMB. Following the common practice in machine learning, we used 80% of the datasets for training and the rest for test. Hyper-parameters were tuned through grid search over a parameter grid and with 5-fold cross-validation to avoid overfitting. The area under the receiver operating characteristic curve (AUC) was used as a model performance measure during the grid search. At the end of the grid search, a model was fitted on the entire training set with the best combination of hyperparameter values. The process of model construction is illustrated in Figure 4. Tree-based machine learning algorithms such as LGMB are not sensitive to feature scale nor the dependency among features. Hence, we constructed seven LGBM models with or without feature scaling and ICA. In contrast, LR models are sensitive to both feature scaling and multicollinearity. Hence, we constructed three RF models with feature scaling and ICA. SVM models are sensitive to feature scaling but less sensitive to multicollinearity. Hence, we constructed six SVM models with feature scaling and with or without ICA. The models were denoted using the following naming rule: [machine learning technique]_[scaling method]_[ICA or none]. For example, rSVM_SS and rSVM_SS_ICA both refer to RBF kernel-based SVM classifiers trained with features that

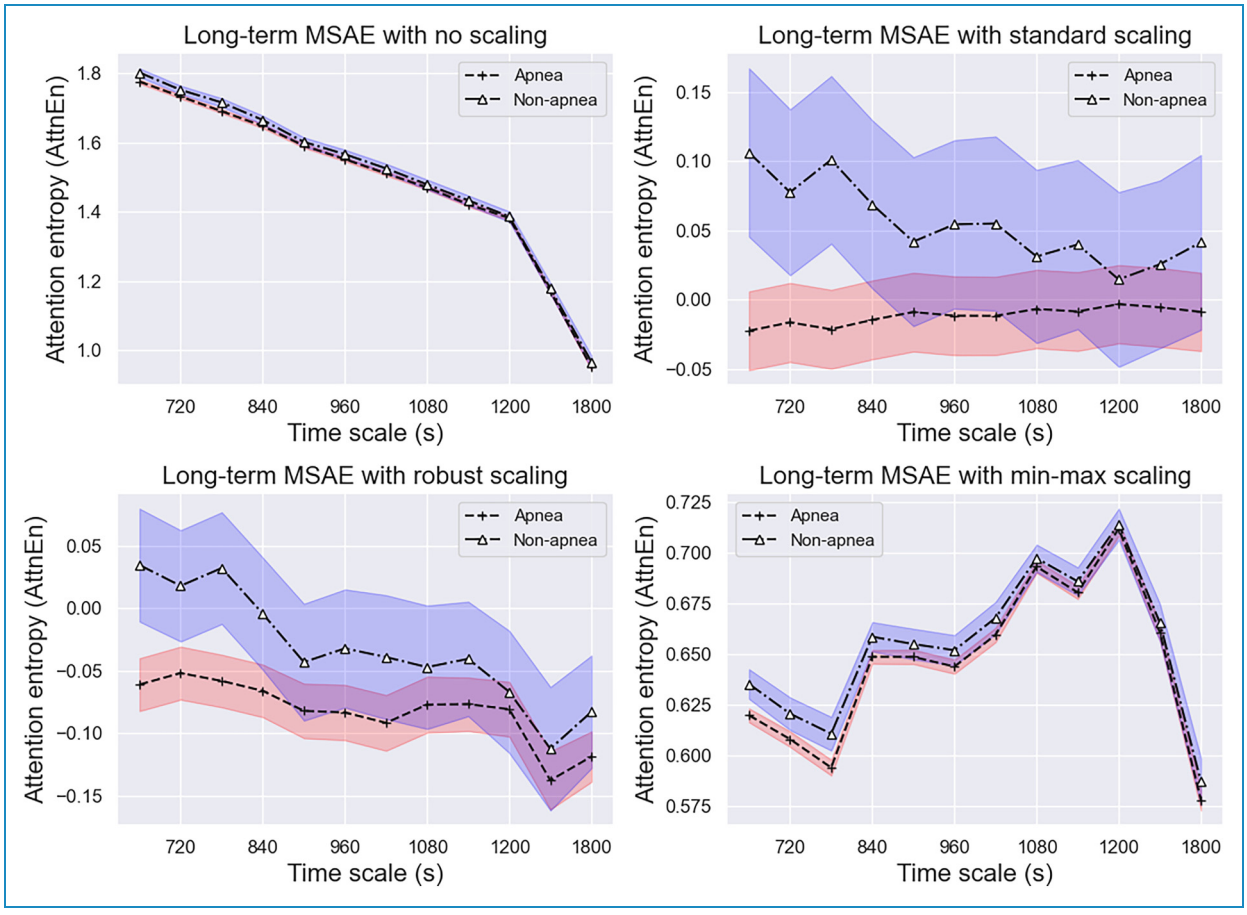

Figure 7. Multiscale attention entropy (MSAE) for long-term scale factor (cutoff ≥ 5/hours).

were normalized using standard scaling, but in the former case, the features were not transformed using ICA while in the latter case, the features were transformed.

As shown in Table 1, there are more samples of the positive class than the negative class, implying an imbalanced dataset. To address this problem, the models (for LR and SVM) altered the loss function by weighting the loss of each sample of its class weight, or (for LGBM) re-weighted the splitting criterion, which is inversely proportional to the class frequency in the input data.

Performance evaluation

After hyperparameters tuning the models were evaluated using multiple performance measures. To account for the imbalanced frequencies of each class, we used Matthew's correlation coefficient (MCC) to evaluate the overall performance of the models. MCC measures the performance of a classification model by summarizing the confusion matrix using equation (3), where TP, TN, FP, and FN denotes true positive, true negative, false positive, and false negative, respectively. The range of the MCC is between −1 (worst) and 1 (best), and 0 corresponds to a prediction made by random guess. It is considered a better performance measure than accuracy (ACC), area under the curve (AUC), and the $F1$-score for imbalanced classification.[47,46,45] Nonetheless, we still calculated ACC, AUC, and $F1$-score to facilitate comparison with prior studies. Sensitivity (SEN) and specificity (SPE) were calculated to quantify the percentage of successfully classified positive cases or negative cases. Positive predictive value (PPV) and negative predictive value (NPV) were also calculated due to their clinical relevance.

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \tag{3}$$

Python 3.10.5 was used for data analysis and for creating the plots. Several Python modules were used, including NumPy, SciPy, scikit-learn, Matplotlib, MNE-Python, and pandas.

Results

Visualization of MSAE

The temporal patterns of attention entropy for different time scales are shown in Figures 5 to 7 for cutoff ≥ 5/hours and

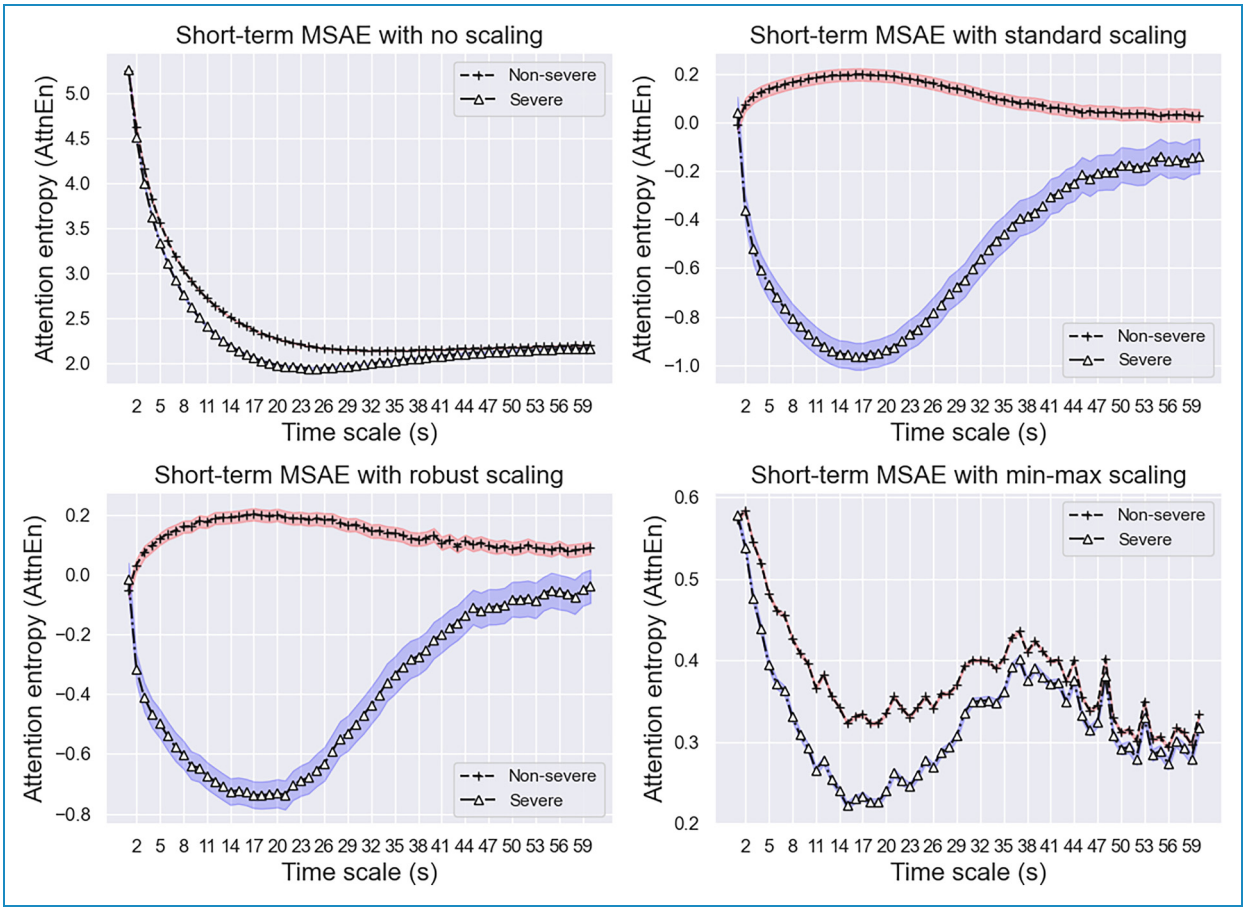

Figure 8. Multiscale attention entropy (MSAE) for short-term scale factor (cutoff ≥ 30/hours).

in Figures 8 to 10 for cutoff ≥ 30/hours. The shaded areas represent the 95% confidence interval for the MSAE curves. Similar trends were observed for both cutoff thresholds. Irrelevant to the scaling method and the cutoff threshold, the MSAE curves of the positive and negative groups exhibit significant differences for short-term and medium-term scale factors. It is shown that the attention entropy of the negative group was consistently higher than that of the positive group when the scale factor was below 600 (i.e. 10 minutes). The difference peaked between scale factors 15 and 20 for both cutoff thresholds. The MSAE curves of the two groups started to converge as the scale factor reached 840 (i.e. 15 minutes) and eventually became inseparable when the scale factor reached 1800 (i.e. 30 minutes). Standard scaling and robust scaling both magnified the between-group differences.

Model performance

As shown in Table 2, for cutoff ≥ 5/hours, the highest MCC (0.402) and AUC (0.747) were achieved by the LR_SS_ICA model. Most of the SVM models and other LR models had similar performance in terms of MCC and AUC, while LGMB models had the poorest performance. Different scaling methods and the ICA process did not significantly affect model performance on the test set. LGMB models were more sensitive but less specific than SVM and LR models, and they also achieved a better trade-off between PPV and NPV. In contrast, SVM and LR models achieved a better tradeoff between SEN and SPE, but at the cost of deteriorated NPV, which implies a decrease in false positive but an increased in false negative.

For cutoff ≥ 30/hours, the highest MCC (0.545) and second highest AUC (0.823) was achieved by the RBF-kernel based SVM models with ICA irrelevant to which scaling method was applied (i.e. rSVM_SS_ICA, rSVM_RS_ICA, and rSVM_MS_ICA). Other SVM models had similar performance, as shown in Table 3. LR models had slightly worse MCC compared to the SVM models, while LGMB models were the weakest in terms of AUC. The ICA process improved the SPE and PPV of the SVM models as well as the AUC, ACC, *F*1, and SEN of the LGMB models. Different scaling methods did not significantly affect model performance. Compared to the SVM and RL models, the LGMB models were less effective in detecting the positive cases but more effective in

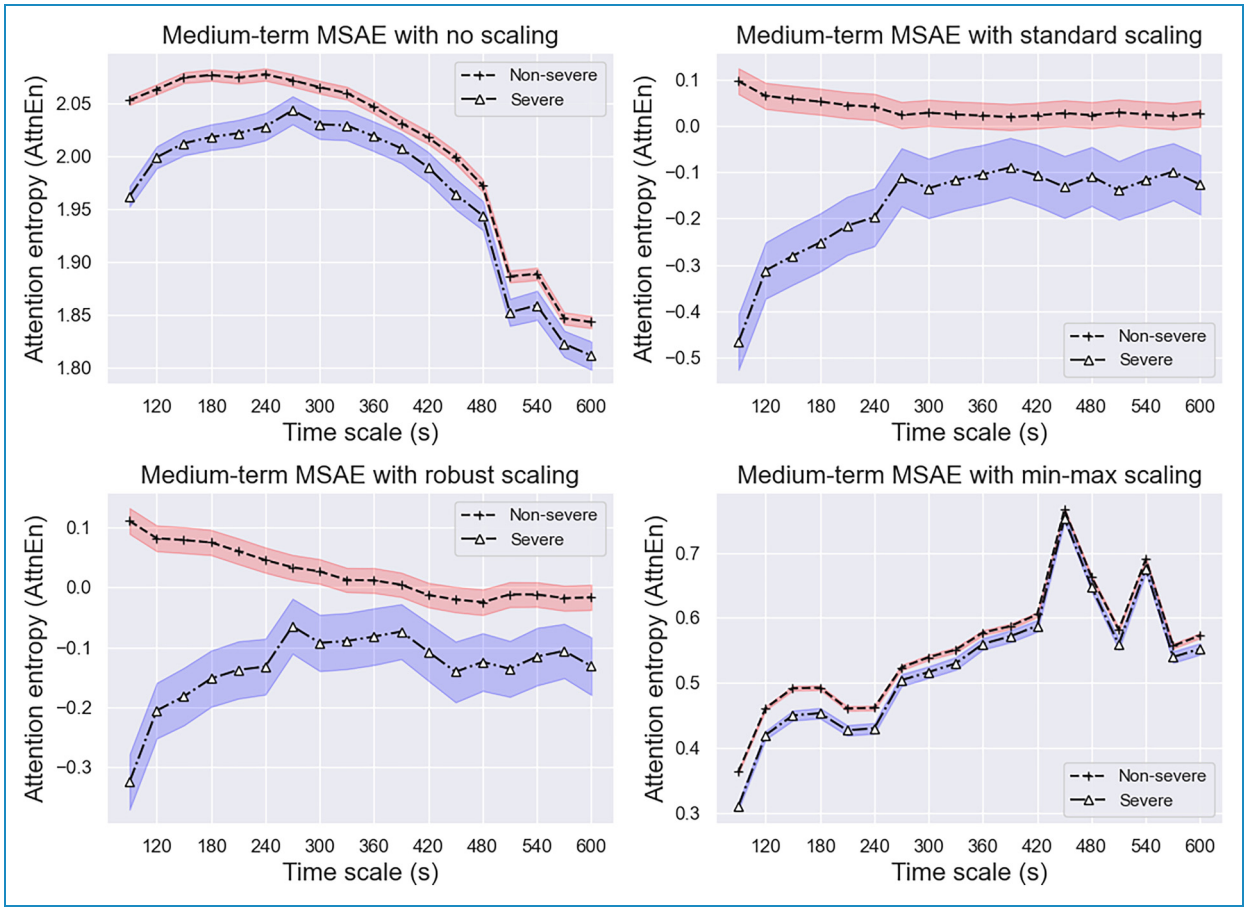

Figure 9. Multiscale attention entropy (MSAE) for medium-term scale factor (cutoff ≥ 30/hours).

detecting the negative cases. The SVM and LR models achieved a good tradeoff between SEN and SPE but at the cost of significantly deteriorated PPV, which implies an increase in false positive rate. In contrast, the LGMB models achieved better tradeoff between PPV and NPV but at the cost of deteriorated SEN, which implies an increase in false negative.

Discussion

Principal results

We have proposed a novel method for easy sleep apnea screening and have presented the evaluated results. In what follows, we discuss the principal findings and their interpretations within the current landscape of the related research field.

To the best of our knowledge, this is the first study that examines multiscale entropy for a wide range of time scales. While prior studies have used single-scale entropy for a scale factor of 1 as a feature to develop apnea screening models,[23,37] our results suggest that single-scale entropy may not have strong distinguishing power between apnea positive and negative groups. As shown in Figures 5 and 8, the attention entropy at a scale factor of 1 of the positive and negative groups are statistically indistinguishable and thus the single-scale entropy measure would have failed in screening the positive cases. As the scale factor increases, however, the between-group difference started to manifest, thus justifying the advantage of using multiscale entropy over single-scale entropy. We found that the MSAE was useful in distinguishing the positive and negative groups at both cutoff thresholds for short- and medium-term time scales. Standard scaling and robust scaling further magnified the differences between groups. The MSAE of the positive group was consistently lower than that of the negative group over a range of time scales for both cutoff thresholds, indicating a loss of complexity in the overnight SpO2 recordings for the positive group. This echoes findings in previous studies that reduced entropy was often observed in the physiological signals recorded from disease populations,[32,44] indicating a loss of complexity with disease. Significant differences were observed between the curves of the two groups when the time scale was below 10 minutes. Notably, the difference between groups peaked around scale factors 15–20 for both cutoffs. As the time scale increases above 15

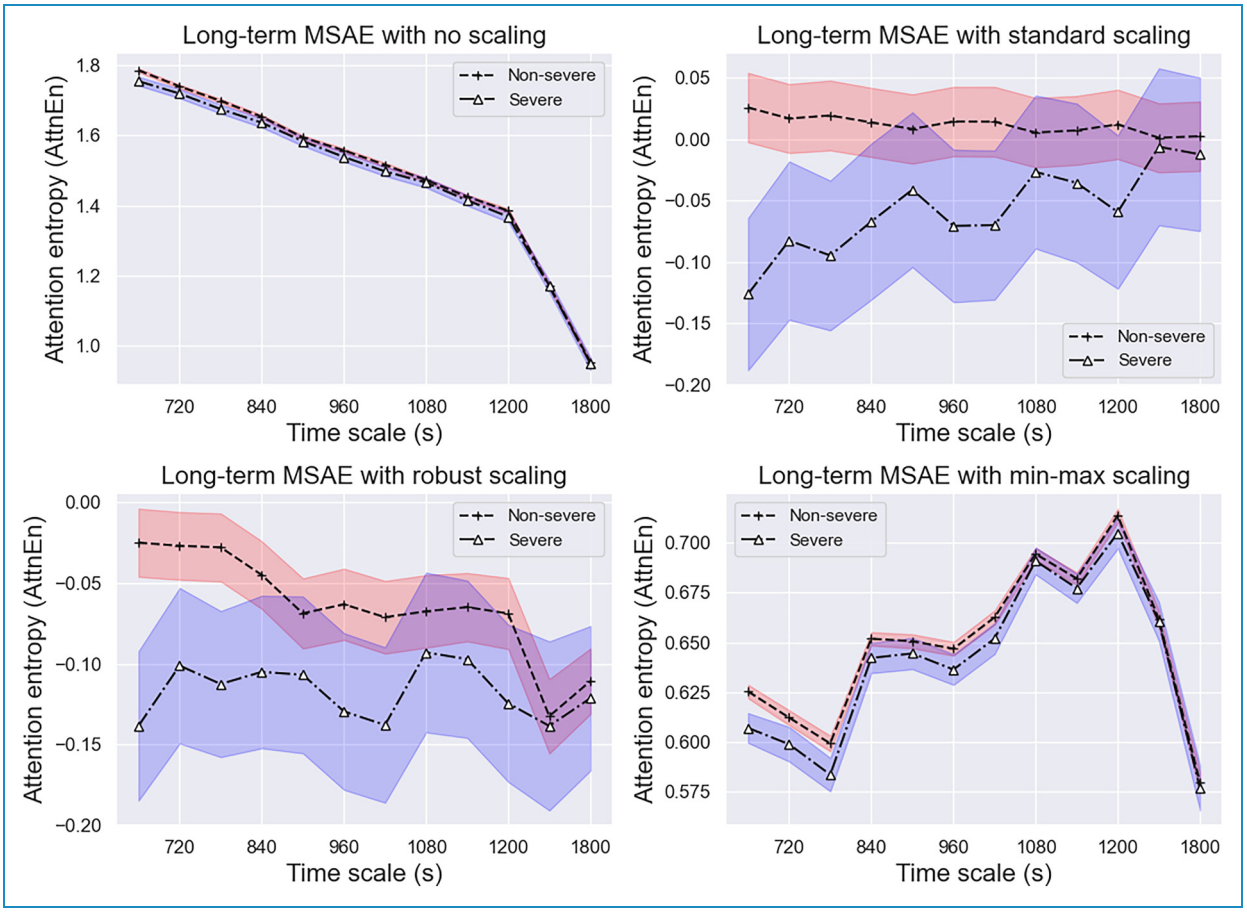

Figure 10. Multiscale attention entropy (MSAE) for long-term scale factor (cutoff ≥ 30/hours).

minutes, the two curves started to overlap and eventually converged when the time scale reached 30 minutes.

Regarding the apnea screening models, the best MCC was 0.402 for cutoff ≥ 5/hours and 0.545 for ≥ 30/hours, respectively, indicating strong positive relationship between the prediction and the ground truth. In addition, the best AUC was 0.747 and 0.823, and the best F1-score was 0.905 and 0.625. Most of the LR and SVM models had similar performances. While many previous studies on apnea detection relied on AUC and F1-score for overall model performance evaluation, we favored MCC over the two as it is more suited for imbalanced classification.[46,47] Moreover, the F1-score is not a useful measure for ≥ 5/hours because, in this case, the negative group is the rare class, whereas in conventional clinical applications, the positive group is usually treated as the rare class. Judged by MCC, our models achieved satisfactory performance, especially for cutoff ≥ 30/hours. We noticed the models had better performance for cutoff ≥ 30/hours than for ≥ 5/hours. One possible explanation could be the distribution of the true AHI at the classification border. As shown in Figure 11, the AHI for both groups near the classification border has similar frequency for cutoff ≥ 5/hours, making it harder to differentiate the two groups. In comparison, the frequency of the two groups at the classification border was more distinguishable for cutoff ≥ 30/hours, as shown in Figure 12.

We also observed that the LGMB models had distinct characteristics compared to the SVM and LR models. They were more effective in detecting the majority class, as indicated by the high SEN for cutoff ≥ 5/hours and high SPE for cutoff ≥ 30/hours. However, their superior performance on the majority class was achieved at the sacrifice of the performance on the minority class. In comparison, the SVM and LR models achieved a better tradeoff between SEN and SPE.

In addition to SEN and SPE, we evaluated the models using PPV and NPV—two relevant measures in clinical settings.[48,49] In this evaluation dimension, the LGMB models achieved a better balance between PPV and NPV, especially for cutoff ≥ 30/hours. The SVM and LR models provided high PPV but low NPV for ≥ 5/hours and the opposite for ≥ 30/hours. The tradeoff between SEN and SPE, and that between PPV and NPV, boils down to the tradeoff between false positive and false negative, where the increase of one is often accompanied by the decrease of the other. Previous studies posit that a high PPV is desirable when the treatment cost is high in relative to its

Table 2. Model performance for cutoff ≥ 5/hours.

	MCC	AUC	ACC (%)	F1-score	SEN (%)	SPE (%)	PPV (%)	NPV (%)
lSVM_SS[a]	0.380	0.734	73.98	0.824	74.30	72.55	92.47	38.34
lSVM_RS	0.379	0.734	73.89	0.823	74.19	72.55	92.46	38.24
lSVM_MS	0.363	0.724	73.27	0.819	73.76	71.08	92.05	37.37
lSVM_SS_ICA	0.369	0.727	73.72	0.822	74.30	71.08	92.10	37.86
lSVM_RS_ICA	0.365	0.725	73.45	0.820	73.97	71.08	92.07	37.56
lSVM_MS_ICA	0.370	0.727	73.81	0.823	74.41	71.08	92.11	37.96
rSVM_SS	0.401	0.746	75.04	0.832	75.27	74.02	92.93	39.74
rSVM_RS	0.383	0.746	75.84	0.840	77.43	68.63	91.81	40.11
rSVM_MS	0.327	0.697	73.54	0.824	75.70	63.73	90.45	36.62
rSVM_SS_ICA	0.374	0.730	74.16	0.826	74.84	71.08	92.15	38.36
rSVM_RS_ICA	0.382	0.734	74.34	0.827	74.84	72.06	92.40	38.68
rSVM_MS_ICA	0.373	0.729	74.07	0.825	74.73	71.08	92.14	38.26
LR_SS_ICA	0.402	0.747	75.13	0.832	75.38	74.02	92.94	39.84
LR_RS_ICA	0.393	0.741	74.78	0.830	75.16	73.04	92.68	39.31
LR_MS_ICA	0.395	0.743	74.78	0.830	75.05	73.53	92.79	39.37
LGMB	0.291	0.594	83.27	0.905	96.76	22.06	84.93	60.00
LGMB_SS	0.270	0.586	82.92	0.903	96.65	20.59	84.67	57.53
LGMB_RS	0.316	0.632	82.30	0.896	93.09	33.33	86.37	51.52
LGMB_MS	0.254	0.575	82.92	0.903	97.30	17.65	84.28	59.02
LGMB_SS_ICA	0.204	0.564	81.77	0.896	96.11	16.67	83.96	48.57
LGMB_RS_ICA	0.284	0.602	82.65	0.900	95.36	25.00	85.23	54.26
LGMB_MS_ICA	0.310	0.614	83.10	0.902	95.36	27.45	85.65	56.57

[a]Model naming rule: [machine learning technique]_[scaling method]_[ICA or none].
lSVM: support vector machine with linear kernel; rSVM: support vector machine with radial basis function kernel; LR: linear regression; LGMB: light gradient boosting machine; SS: standard scaling; RS: robust scaling; MS: min-max scaling; SEN: sensitivity; SPE: specificity; PPV: positive predictive value; NPV: negative predictive value.

potential benefits, and a high NPV is desirable when the disease condition is serious, contagious, or likely to progress quickly.[49] Using those rules of thumb, the models that provided a high PPV and a moderate NPV are desirable for our problem of interest, as sleep apnea is not an urgent or contagious condition and the treatment cost (e.g. CPAP and surgery) may outweigh its benefit. To this end, the LR_SS_ICA model and the LGMB_MS_ICA model may be the most suited model for each cutoff threshold if evaluated along this dimension.

Limitations

The present study has several limitations that are amenable to future investigations. First, the dataset used for model

Table 3. Model performance for cutoff ≥ 30/hours.

	MCC	AUC	ACC (%)	F1-score	SEN (%)	SPE (%)	PPV (%)	NPV (%)
lSVM_SS[a]	0.518	0.814	80.44	0.600	83.00	79.89	47.03	95.62
lSVM_RS	0.524	0.819	80.53	0.604	84.00	79.78	47.19	95.87
lSVM_MS	0.524	0.819	80.53	0.604	84.00	79.78	47.19	95.87
lSVM_SS_ICA	0.526	0.818	80.97	0.607	83.00	80.54	47.84	95.66
lSVM_RS_ICA	0.526	0.818	80.97	0.607	83.00	80.54	47.84	95.66
lSVM_MS_ICA	0.522	0.815	80.88	0.604	82.50	80.54	47.69	95.54
rSVM_SS	0.533	0.824	80.80	0.610	85.00	79.89	47.62	96.12
rSVM_RS	0.518	0.817	79.91	0.598	84.50	78.92	46.30	95.95
rSVM_MS	0.477	0.787	79.47	0.572	77.50	79.89	45.32	94.29
rSVM_SS_ICA	0.545	0.823	82.57	0.625	82.00	82.69	50.46	95.53
rSVM_RS_ICA	0.545	0.823	82.57	0.625	82.00	82.69	50.46	95.53
rSVM_MS_ICA	0.545	0.823	82.57	0.625	82.00	82.69	50.46	95.53
LR_SS_ICA	0.511	0.808	80.62	0.597	81.00	80.54	47.23	95.17
LR_RS_ICA	0.511	0.808	80.62	0.597	81.00	80.54	47.23	95.17
LR_MS_ICA	0.511	0.808	80.62	0.597	81.00	80.54	47.23	95.17
LGMB	0.529	0.714	87.88	0.573	46.00	96.88	76.03	89.30
LGMB_SS	0.509	0.699	87.52	0.547	42.50	97.20	76.58	88.71
LGMB_RS	0.466	0.686	86.46	0.517	41.00	96.24	70.09	88.35
LGMB_MS	0.525	0.717	87.70	0.575	47.00	96.45	74.02	89.43
LGMB_SS_ICA	0.518	0.714	87.52	0.569	46.50	96.34	73.23	89.33
LGMB_RS_ICA	0.506	0.712	87.17	0.562	46.50	95.91	70.99	89.29
LGMB_MS_ICA	0.539	0.716	88.14	0.579	46.00	97.20	77.97	89.33

[a]Model naming rule: [machine learning technique]_[scaling method]_[ICA or none].
lSVM: support vector machine with linear kernel; rSVM: support vector machine with radial basis function kernel; LR: linear regression; LGMB: light gradient boosting machine; SS: standard scaling; RS: robust scaling; MS: min–max scaling; SEN: sensitivity; SPE: specificity; PPV: positive predictive value; NPV: negative predictive value.

construction has a high prevalence of sleep apnea. The model performance, especially the measures that are heavily affected by disease prevalence including PPV and NPV, should be interpreted with caution and may not be generalized to other datasets. Second, we did not consider subtypes of sleep apnea and comorbidities in this study. Third, we did not use conventional time-domain and frequency-domain features that are commonly used in machine learning-based sleep apnea detection. Further studies may leverage those features to improve model performance.

Comparison with prior work

Many systems and algorithms have been developed for home sleep apnea screening, but most prior studies have only been validated on small samples (N = 3–481),[50]

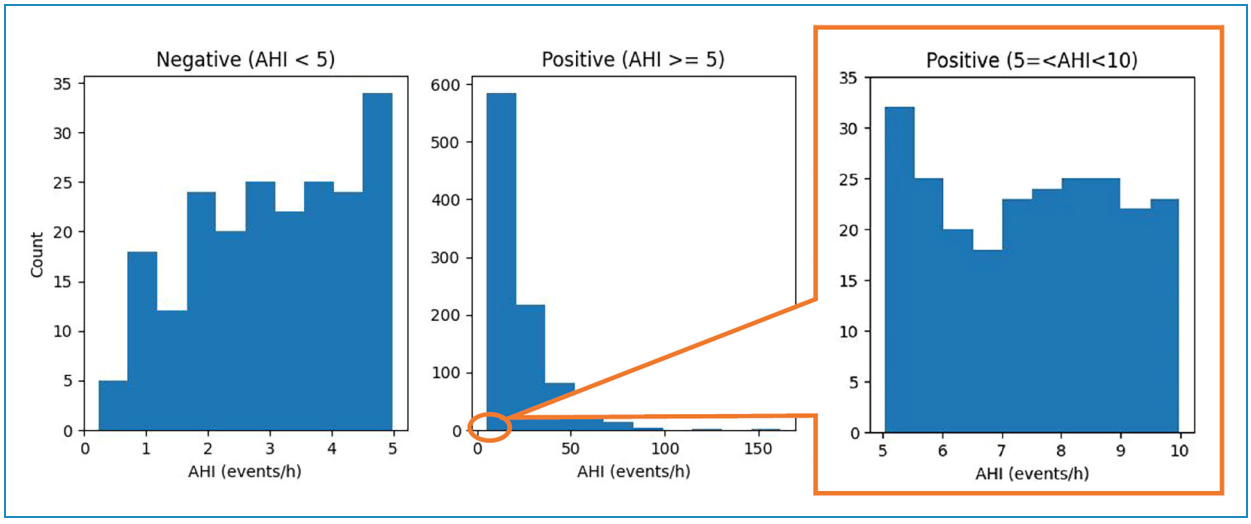

Figure 11. Distribution of apnea–hypopnea index (AHI) for the positive and negative groups (cutoff ≥ 5/hours).

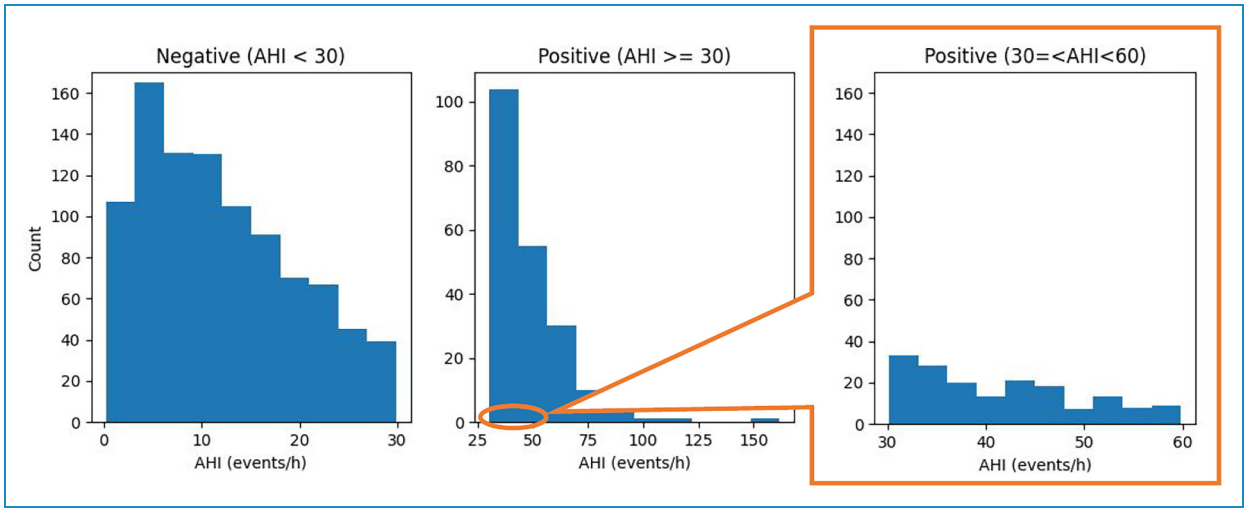

Figure 12. Distribution of apnea–hypopnea index (AHI) for the positive and negative class at cutoff ≥ 30/hours.

which threats the rigidity and generalizability of the developed models. To the best of our knowledge, there are only few studies that have validated their methods on a sufficiently large sample ($N > 1000$) and have used cutoff thresholds that permit direct comparisons with our models.[9,12,8,51] The study that has validated their models on the largest sample so far ($N = 17{,}448$) used a different cutoff threshold of ≥ 15/hours[51] and thus forbids a direct comparison to our study. That model only relies on four simple demographic features: age, sex, BMI, race, and achieved an AUC between 0.61 and 0.72, a high sensitivity between 0.65 and 0.91 but a low specificity between 0.36 and 0.53. The deep learning-based sleep apnea screening model OxiNet was also validated on a large sample ($N = 12{,}923$) and achieved an average F1-score between 0.75 and 0.84.[25] Similar to the present study, the OxiNet model only relies on single channel overnight SpO2 signals. However, it was intended for four-class classification rather than binary classification and thus also forbids a direct comparison to our study.

Table 4 presents the comparison between our models and existing models for cutoff ≥ 5/hours and ≥ 30/hours, respectively. Our models achieved comparable performance to the best existing model (i.e. the SVM model by Huang et al.[9] for ≥ 5/hours), and better performance for ≥ 30/hours. Among the three studies that allow direct comparison with our study, only one study Huang et al.[9] has used a sample larger than our study. Furthermore, models developed in those studies rely on self-reported symptoms[8,12] and/or complicated laboratory blood reports.[8]

Table 4. Comparison to prior studies.

	Model	ACC (%)	AUC	F1-score	SEN (%)	SPE (%)	PPV (%)	NPV (%)
Cutoff ≥ 5/hours								
	SVM[8]	68.06	0.65	0.76	88.76	40.74	66.36	73.33
	SVM[9]	74.24	0.82	0.83	74.14	74.71	93.23	38.15
	LR[9]	73.77	0.84	–	94.41	37.87	72.55	79.56
	BQ[9]	67.58	0.54	–	74.95	32.91	84.01	21.89
	NoSAS Score[9]	57.25	0.70	–	50.62	88.39	95.31	27.58
	SLIM (10 size)[9]	54.68	0.69	0.63	47.10	90.30	95.80	26.64
	SLIM (10 size)[12]	–	0.79	–	64.20	77.00	–	–
	STOP-Bang[12]	–	–	–	83.60	56.40	–	–
	Our model (rSVM_SS_ICA)	blue74.16	blue0.73	blue0.83	blue74.84	blue71.08	blue92.15	blue38.36
	Our model (LR_SS_ICA)	blue75.13	blue0.75	blue0.83	blue75.38	blue74.02	blue92.94	blue39.84
	Our model (LGMB_RS)	blue82.30	blue0.63	blue0.90	blue93.09	blue33.33	blue86.37	blue51.52
Cutoff ≥ 30/hours								
	SVM[9]	70.28	0.78	0.66	70.26	70.30	61.93	77.86
	LR[9]	72.83	0.79	–	65.01	78.77	69.94	74.77
	BQ[9]	48.09	0.53	–	76.68	28.55	42.31	64.17
	NoSAS Score[9]	68.30	0.68	–	64.88	70.64	60.16	74.64
	SLIM (10 size)[9]	69.40	0.68	0.62	62.24	74.29	62.33	74.22
	Our model (rSVM_SS_ICA)	82.57	0.82	0.63	82.00	82.69	50.46	95.53
	Our model (LR_SS_ICA)	80.62	0.81	0.60	81.00	80.54	47.23	95.17
	Our model (LGMB_MS_ICA)	88.14	0.72	0.58	46.00	97.20	77.97	89.33

SVM: support vector machine; LR: logistic regression; NoSAS: Neck circumference, Obesity, Snoring, Age, Sex; SLIM: supersparse linear integer model; SEN: sensitivity; SPE: specificity; PPV: positive predictive value; NPV: negative predictive value.

Those models are thus not useful for patients who have no self-aware symptoms or for cases where the required medical information is not available. Another pitfall of the prior studies is that the models often yield a high sensitivity at the cost of a low specificity, whereas our models, especially the SVM and LR based ones, achieved a good tradeoff between sensitivity and specificity. Similar to the dataset used in this study, those used by Huang et al.[9] and Ustun et al.[12] also had the problem of class imbalance with the positive class significantly outnumbering the negative class. However, neither of the two studies explicitly addressed the class imbalance issue in model training and evaluation. In comparison, our model training explicitly accounted for class imbalance, and thus could potentially outperform existing models even if more appropriate measures such as MCC was used. In addition, our models do not rely on self-reported symptoms and thus can be useful even for asymptomatic patients.

Our method has several fundamental advantages over existing models. Compared to models that use traditional features, the MSAE features of overnight SpO2 are easier to obtain. EHR features such as medication records and

blood test results are not always available, and questionnaire scores cannot be calculated from asymptomatic patients. Conventional time-domain, frequency-domain, and single-scale nonlinear features do not have satisfying discriminating power as indicated by the sub-optimal performance of prior shallow learning models. As shown in Figures 5 to 10, the fluctuation of the scaled attention entropy demonstrates the significant difference in the positive and negative groups at various time scales, which implies that the MSAE features are potentially powerful features for classification. In addition, this is the first study that investigates the MSE of the overnight SpO2 signals at a wide range of time scales. While prior studies on the MSE of physiological signals only used time scales lower than 1 minute, our study investigated a wide range of time scales between 1 seconds and 30 minutes. Our analysis found that the entropy at a longer time scale (1–15 minutes) could still be useful features for distinguishing apnea positive and negative. This is a new finding that no prior study had discovered. The advantage of our approach over deep learning models lies in its simplicity and better explainability. For one thing, shallow learning models such as ours do not require heavy computational and storage resources and can be easily applied to structured data. For another, our models have better explainability because the features used to develop the models have a physical meaning. In particular, the MSAE provides insights into how the complexity of the SpO2 signals changes at different time scales. Figures 5 to 10 show a loss of complexity for sleep apnea indicated by lower attention entropy values over a range of time scales.

Conclusions

In this study, we proposed a novel method for sleep apnea screening using only overnight SpO2 signals and simple demographic information. The method computes the attention entropy of the SpO2 signals for different time scale and uses these MSAE, together with age, sex, BMI, and blood pressure, as features to construct classification models that automatically detect positive cases for cutoff thresholds $\geq$ 5/hours and $\geq$ 30/hours. Depending on the machine learning algorithm adopted, feature scaling and/or ICA were applied to the original feature set so that the transformed feature set met the assumptions of the algorithms. The best models achieved an MCC of 0.402 for cutoff $\geq$ 5/hours and 0.545 for $\geq$ 30/hours, respectively, indicating a strong positive relationship between the prediction and the ground truth. Compared to prior studies, our models achieved comparable or better performance and have the merit of not relying on self-reported symptoms and thus can be useful even for asymptomatic patients.

Acknowledgements: The author would like to thank the National Sleep Research Resource for sharing the SHHS dataset.

Contributorship: ZL conceived of the presented idea, performed the computations, verified the analytical methods, interpreted the results, and wrote the final manuscript.

Declarations of conflicting interest: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article. The author(s) declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.

Ethical approval: Ethics approval was obtained from the Ethics Review Board at Kyoto University of Advanced Science. Access to the SHHS database was granted by the National Sleep Research Resource.

Funding: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was supported by the JSPS KAKENHI Grant Number 21K17670.

Guarantor: ZL.

ORCID iD: Zilu Liang https://orcid.org/0000-0002-2328-5016

References

1. Fujita Y and Yamauchi M. Diagnosis of sleep apnea [in Japanese]. *J Jap Soc Interna Med* 2020; 109: 6.
2. Benjafield AV, Ayas NT, Eastwood PR et al. Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis. *Lancet Respir Med* 2019; 7: 687–698.
3. Knauert M, Naik S, Gillespie MB, et al. Clinical consequences and economic costs of untreated obstructive sleep apnea syndrome. *World J Otorhinolaryngol Head Neck Surg* 2015; 1: 17–27.
4. Bhattacharjee A, Saha S, Fattah SA et al. Sleep apnea detection based on Rician modeling of feature variation in multiband EEG signal. *IEEE J Biomed Health Inform* 2019; 23: 1066–1074.
5. Bahrami M and Forouzanfar M. Sleep apnea detection from single-lead ECG: a comprehensive analysis of machine learning and deep learning algorithms. *IEEE Trans Instrum Meas* 2022; 71: 1–11.
6. Han H and Oh J. Application of various machine learning techniques to predict obstructive sleep apnea syndrome severity. *Sci Rep* 2023; 13: 6379.
7. Casal-Guisande M, Torres-Durán M, Mosteiro-Añón M, et al. Design and conceptual proposal of an intelligent clinical decision support system for the diagnosis of suspicious obstructive sleep apnea patients from health profile. *Int J Environ Res Public Health* 2023; 20: 3627.
8. Ramesh J, Keeran N, Sagahyroon A, et al. Towards validating the effectiveness of obstructive sleep apnea classification from electronic health records using machine learning. *Healthcare* 2021; 9: 1450.

9. Huang WC, Lee PL, Liu YT et al. Support vector machine prediction of obstructive sleep apnea in a large-scale Chinese clinical sample. *Sleep* 2020; 43: 1–11.
10. Ferreira-Santos D and Rodrigues PP. A clinical risk matrix for obstructive sleep apnea using Bayesian network approaches. *Int J Data Sci Anal* 2019; 8: 339–349.
11. Zoroglu C and Turkeli S. Fuzzy expert system for severity prediction of obstructive sleep apnea hypopnea syndrome. *The Journal of Cognitive Systems* 2017; 2: 37–43.
12. Ustun B, Westover MB, Rudin C et al. Clinical prediction models for sleep apnea: the importance of medical history over symptoms. *J Clin Sleep Med* 2016; 12: 161–168.
13. Sun LM, Chiu HW, Chuang CY, et al. A prediction model based on an artificial intelligence system for moderate to severe obstructive sleep apnea. *Sleep Breath* 2011; 15: 317–323.
14. Liang Z and Chapa-Martell MA. A multi-level classification approach for sleep stage prediction with processed data derived from consumer wearable activity trackers. *Frontiers in Digital Health* 2021; 3: 66594.
15. Liang Z, Ploderer B, Martell MAC, et al. A cloud-based intelligent computing system for contextual exploration on personal sleep-tracking data using association rule mining. In: Martin-Gonzalez A and Uc-Cetina V (eds) *Intelligent computing systems. ISICS 2016. Communications in computer and information science*. Cham: Springer, 2016.
16. Walch O, Huang Y, Forger D et al. Sleep stage prediction with raw acceleration and photoplethysmography heart rate data derived from a consumer wearable device. *Sleep* 2019; 42: zsz180.
17. Xie B and Minn H. Real-time sleep apnea detection by classifier combination. *IEEE Trans Inf Technol Biomed* 2012; 16: 469–477.
18. Deviaene M, Lázaro J, Huysmans D et al. Sleep apnea detection using pulse photoplethysmography. In *In Proceedings of the 2018 Computing in Cardiology Conference (CinC)*. IEEE.
19. Mostafa SS, Mendonça F, Morgado-Dias F et al. SpO2 based sleep apnea detection using deep learning. In *In Proceedings of the 2017 IEEE 21st International Conference on Intelligent Engineering Systems (INES)*.
20. Ravelo-García AG, Kraemer JF, Navarro-Mesa JL et al. Oxygen saturation and RR intervals feature selection for sleep apnea detection. *Entropy* 2015; 5: 2932–2957.
21. Almazaydeh L, Faezipour M and Elleithy K. A neural network system for detection of obstructive sleep apnea through SpO2 signal features. *International Journal of Advanced Computer Science and Applications(IJACSA)* 2012; 3: 7–11.
22. Behar JA, Palmius N, Li Q, et al. Feasibility of single channel oximetry for mass screening of obstructive sleep apnea. *EClinicalMedicine* 2019; 11: 81–88.
23. Gutierrez-Tobal GC, Alvarez D, Crespo A et al. Evaluation of machine-learning approaches to estimate sleep apnea severity from at-home oximetry recordings. *IEEE J Biomed Health Inform* 2019; 23: 882–892.
24. Sharma P, Jalali A, Majmudar M, et al. Deep-learning based sleep apnea detection using SpO2 and pulse rate. *Annu Int Conf IEEE Eng Med Biol Soc* 2022; 2022: 2611–2614.
25. Levy J, Alvarez D, Del Campo F, et al. Deep learning for obstructive sleep apnea diagnosis based on single channel oximetry. *Nat Commun* 2023; 14: 4881.
26. Mencar C, Gallo C, Mantero M et al. Application of machine learning to predict obstructive sleep apnea syndrome severity. *Health Informatics J* 2020; 26: 298–317.
27. Quan SF, Howard BV, Iber C et al. The sleep heart health study: design, rationale, and methods. *Sleep* 1997; 20: 1077–1085.
28. Zhang GQ, Cui L, Mueller R et al. The national sleep research resource: towards a sleep data commons. *J Am Med Inform Assoc* 2018; 25: 1351–1358.
29. Bernardini A, Brunello A, Gigli GL et al. Osasud: a dataset of stroke unit recordings for the detection of obstructive sleep apnea syndrome. *Sci Data* 2022; 9: 177.
30. Magalang UJ, Dmochowski J, Veeramachaneni S et al. Prediction of the apnea–hypopnea index from overnight pulse oximetry. *Chest* 2003; 124: 1694–1701.
31. Berry RB, Brooks R, Gamaldo C et al. *The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications, Version 2.4*. Darien, IL: American Academy of Sleep Medicine, 2017.
32. Madalena C, Goldberger AL and K PC. Multiscale entropy analysis of complex physiologic time series. *Phys Rev Lett* 2002; 89: 068102.
33. Guo Y, Chen Y, Yang Q et al. Multi-scale permutation entropy: a potential measure for the impact of sleep medication on brain dynamics of patients with insomnia. *Entropy* 2021; 23: 1101
34. Crespo A, Álvarez D, SNMCGutiérrez-Tobal G et al. Multiscale entropy analysis of unattended oximetric recordings to assist in the screening of paediatric sleep apnoea at home. *Entropy* 2017; 19: 284.
35. Pan WY, Su MC, Wu HT et al. Multiscale entropic assessment of autonomic dysfunction in patients with obstructive sleep apnea and therapeutic impact of continuous positive airway pressure treatment. *Sleep Med* 2016; 20: 12–17.
36. Yang J, Choudhary GI, Rahardja S et al. Classification of interbeat interval time-series using attention entropy. *IEEE Trans Affect Comput* 2020; 14: 321–330.
37. Hornero R, Alvarez D and del Campo DSEP, Abásolo F and Zamarron C. Utility of approximate entropy from overnight pulse oximetry data in the diagnosis of the obstructive sleep apnea syndrome. *IEEE Trans Biomed Eng* 2007; 54: 107–113.
38. Fietze I, Laharnar N, Obst A et al. Prevalence and association analysis of obstructive sleep apnea with gender and age differences—results of ship-trend. *J Sleep Res* 2019; 28: e12770.
39. Marrone O and Bonsignore MR. Blood-pressure variability in patients with obstructive sleep apnea: current perspectives. *Nat Sci Sleep* 2018; 10: 229–242.
40. Bell AJ and Sejnowski TJ. An information-maximization approach to blind separation and blind deconvolution. *Neural Comput* 1995; 7: 1129–1159.
41. Kwak N and Choi CH. Feature extraction based on ICA for binary classification problems. *IEEE Trans Knowl Data Eng* 2003; 15: 1374–1388.
42. Ben-Hur A, Ong CS, Sonnenburg S et al. Support vector machines and kernels for computational biology. *PLoS Comput Biol* 2008; 4: 1–10.
43. Al-Angari HM and Sahakian AV. Automated recognition of obstructive sleep apnea syndrome using support vector machine classifier. *IEEE Trans Inf Technol Biomed* 2012; 16: 463–468.
44. Pan WY, Su MC, Wu HT et al. Multiscale entropy analysis of heart rate variability for assessing the severity of sleep disordered breathing. *Entropy* 2015; 17: 231–243.
45. Chicco D and Jurman G. The Matthew's correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. *BioData Min* 2023; 16: 1.

46. Chicco D, Totsch N and Jurman G. The Matthew's correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData Min* 2021; 14: 13.
47. Chicco D and Jurman G. The advantages of the Matthew's correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* 2020; 21:1–13. Article 6.
48. Akobeng AK. Understanding diagnostic tests 1: sensitivity, specificity and predictive values. *Acta Paediatr* 2007; 96: 338–341.
49. Trevethan R. Sensitivity, specificity, and predictive values: foundations, pliabilities, and pitfalls in research and practice. *Front Public Health* 2017; 5: 307.
50. Mendonça F, Mostafa SS, Ravelo-García AG et al. Devices for home detection of obstructive sleep apnea: a review. *Sleep Med Rev* 2018; 41: 149–160.
51. Holfinger SJ, Lyons MM, Keenan BT et al. Diagnostic performance of machine learning-derived OSA prediction tools in large clinical and community-based samples. *Chest* 2022; 161: 807–817.