



OPEN ACCESS

EDITED BY

Sy Duong-Quy,
Lam Dong Medical College, Vietnam

REVIEWED BY

Elissa Ye,
Massachusetts General Hospital, Harvard
Medical School, United States
Tilman Huppertz,
Johannes Gutenberg University Mainz,
Germany
Sami Myllymaa,
University of Eastern Finland, Finland

*CORRESPONDENCE

Hyang Woon Lee
✉ leeh@ewha.ac.kr

[†]These authors have contributed equally to this work

RECEIVED 08 October 2022

ACCEPTED 03 May 2023

PUBLISHED 14 June 2023

CITATION

Kang C, An S, Kim HJ, Devi M, Cho A,
Hwang S and Lee HW (2023) Age-integrated
artificial intelligence framework for sleep stage
classification and obstructive sleep apnea
screening.
Front. Neurosci. 17:1059186.
doi: 10.3389/fnins.2023.1059186

COPYRIGHT

© 2023 Kang, An, Kim, Devi, Cho, Hwang and Lee. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Age-integrated artificial intelligence framework for sleep stage classification and obstructive sleep apnea screening

Chaewon Kang^{1†}, Sora An^{2†}, Hyeon Jin Kim^{3,4}, Maithreyee Devi¹,
Aram Cho⁵, Sungeun Hwang⁶ and Hyang Woon Lee^{1,4,7*}

¹Computational Medicine, System Health Science and Engineering Program, Ewha Womans University, Seoul, Republic of Korea, ²Department of Communication Disorders, Ewha Womans University, Seoul, Republic of Korea, ³Department of Neurology, Korea University Ansan Hospital, Ansan, Republic of Korea, ⁴Department of Neurology, Ewha Womans University School of Medicine, Seoul, Republic of Korea, ⁵Department of Nursing Science, Ewha Womans University, Seoul, Republic of Korea, ⁶Department of Neurology, Ewha Womans University Mogdong Hospital, Seoul, Republic of Korea, ⁷Department of Medical Science, Ewha Womans University School of Medicine and Ewha Medical Research Institute, Seoul, Republic of Korea

Introduction: Sleep is an essential function to sustain a healthy life, and sleep dysfunction can cause various physical and mental issues. In particular, obstructive sleep apnea (OSA) is one of the most common sleep disorders and, if not treated in a timely manner, OSA can lead to critical problems such as hypertension or heart disease.

Methods: The first crucial step in evaluating individuals' quality of sleep and diagnosing sleep disorders is to classify sleep stages using polysomnographic (PSG) data including electroencephalography (EEG). To date, such sleep stage scoring has been mainly performed manually *via* visual inspection by experts, which is not only a time-consuming and laborious process but also may yield subjective results. Therefore, we have developed a computational framework that enables automatic sleep stage classification utilizing the power spectral density (PSD) features of sleep EEG based on three different learning algorithms: support vector machine, k-nearest neighbors, and multilayer perceptron (MLP). In particular, we propose an integrated artificial intelligence (AI) framework to further inform the risk of OSA based on the characteristics in automatically scored sleep stages. Given the previous finding that the characteristics of sleep EEG differ by age group, we employed a strategy of training age-specific models (younger and older groups) and a general model and comparing their performance.

Results: The performance of the younger age-specific group model was similar to that of the general model (and even higher than the general model at certain stages), but the performance of the older age-specific group model was rather low, suggesting that bias in individual variables, such as age bias, should be considered during model training. Our integrated model yielded an accuracy of 73% in sleep stage classification and 73% in OSA screening when MLP algorithm was applied, which indicates that patients with OSA could be screened with the corresponding accuracy level only with sleep EEG without respiration-related measures.

Discussion: The current outcomes demonstrate the feasibility of AI-based computational studies that when combined with advances in wearable devices and relevant technologies could contribute to personalized medicine by not only

assessing an individuals' sleep status conveniently at home but also by alerting them to the risk of sleep disorders and enabling early intervention.

KEYWORDS

polysomnography, electroencephalography, sleep staging, obstructive sleep apnea, age-integrated, machine learning, artificial intelligence

1. Introduction

Sleep is an essential part in human life. Poor sleep quality can lead to reduced physical performance and have a negative impact on cognitive functions (Yuan et al., 2019). The number of patients with sleep disorders has been constantly increasing due to light pollution at night, shift work, and altered lifestyles with the recent pandemic of COVID-19 (Marvaldi et al., 2021). One of the most common sleep disorders is obstructive sleep apnea (OSA), which is estimated to affect one-seventh of the global population (Lyons et al., 2020). Patients with OSA have interrupted sleep because they repeatedly stop and resume breathing while they are asleep. They not only have difficulty with daytime activities but can also develop serious health concerns, such as hypertension and heart problems, if their OSA is not treated in a timely manner (Kumari et al., 2020).

To evaluate sleep quality and diagnose sleep disorders, polysomnography (PSG) is widely used. PSG measures various bio-signals including electroencephalography (EEG), electromyography (EMG), and electrooculography (EOG) signals and respiratory and cardiac activities. Through comprehensive analyses using the acquired multimodal data, individuals' sleep conditions are assessed. A crucial first task in such analyses is to classify sleep stages based on the PSG data (mainly, sleep EEG). To date, sleep scoring has been performed manually by experts following standardized manuals such as the American Academy of Sleep Medicine (AASM) manual (Loh et al., 2020). According to the AASM manual, sleep is classified into the following five stages: wake, rapid eye movement (REM), and three non-REM (NREM) stages including N1, N2, and N3. Experts divide the acquired PSG data into 30-s epochs and then assign sleep stages to each epoch according to standardized criteria. Sleep stage scoring can be a time-consuming, laborious process because it is performed manually and because the time length of one full night of collected PSG data is approximately 7–8 h. Another problem with traditional sleep scoring is that the results might be subjective depending on which experts conducted the scoring. Therefore, it is desirable to devise an automatic process for sleep staging to overcome the abovementioned difficulties.

OSA is diagnosed by comprehensive analyses using questionnaires and various bio-signals obtained during PSG. In particular, airflow *via* a nasal pressure sensor and thermistor, oxygen saturation *via* pulse oximetry, and respiratory effort *via* chest and abdominal belts are used as important indices. EEG, electrocardiography (ECG), EOG, and EMG signals emitted during sleep are also used for diagnosis (Gottlieb and Punjabi, 2020). Based on these bio-signals, indices to evaluate sleep quality, including total sleep time, time spent in each sleep stage, frequency of arousal, and apnea–hypopnea index (AHI, the number

of apnea and hypopnea events per hour of sleep) are calculated, which determine the presence and severity of OSA (Patil et al., 2007). Recent studies have demonstrated that there are significant differences in sleep EEG between OSA and healthy groups (Kumari et al., 2020; Kang et al., 2021), suggesting the need to focus more on sleep EEG. In particular, Kang et al. (2021) demonstrated marked differences in the power spectral densities (PSDs) of beta and sigma frequency bands and indicated that those differences were more pronounced during NREM than during REM stages.

Meanwhile, following the development of artificial intelligence (AI) technologies, recently, the studies attempting to automatically classify the sleep stages based on PSG data are increasing (Zhao et al., 2019; Sekkal et al., 2022). In particular, studies that have built sleep stage scoring models by applying conventional machine learning techniques, such as linear discriminant analysis (Liang et al., 2012; Long et al., 2013), k-nearest neighbors (kNN) (Li et al., 2012), and support vector machine (SVM) (Willemsen et al., 2013; Huang et al., 2014; Wu et al., 2014; Zhu et al., 2014; Acharya et al., 2015; Enshaeifar et al., 2015), and overall, they have shown a classification performance of 70–90%. More recently, studies utilizing deep learning and artificial neural networks have been reported, where they adopted various architectures, including convolutional neural networks (CNNs; Chambon et al., 2018; Mikkelsen and De Vos, 2018; Supratak and Guo, 2020), recurrent neural networks (RNNs; Malafeev et al., 2018; Phan et al., 2018, 2019), deep neural networks (DNNs; Wei et al., 2018), or combinations thereof, such as CNN + RNN (Biswal et al., 2018; Stephansen et al., 2018; Korkalainen et al., 2019; Mousavi et al., 2019) and DNN + RNN (Dong et al., 2018). These state-of-the-art studies have further improved the performance of sleep stage classifications to the 80–90% level, without explicitly defining classification rules or features of each sleep stage. However, depending on the complexity of the model, more computational resources and time are required, thus, the trade-off between model accuracy and computational load may need to be considered, depending on the circumstances (e.g., in a mobile device-dependent environment) (Janiesch et al., 2021).

To date, studies in this area have been mainly conducted based on data obtained from healthy subjects (particularly younger adults). Recently, machine learning studies using large datasets including data from patients with sleep disorders and subjects with wide age ranges have been reported (Korkalainen et al., 2019; Jarchi et al., 2020; Sharma et al., 2021; Hussain et al., 2022). In particular, Korkalainen et al. (2019) proposed a deep learning-based sleep stage classification model using a clinical dataset of patients with suspected OSA and demonstrated that classification accuracy decreased as the OSA severity increased (84.5% for individuals without OSA; 76.5% for severe OSA patients). Sharma et al. built the models by applying six traditional machine learning

classifiers (decision trees, logistic regression, naive Bayes, SVM, kNN, and ensemble bagged trees) based on a dataset, which included healthy subjects and patients with multiple sleep disorders, including insomnia, narcolepsy, REM behavior disorder, etc., and achieved a maximum accuracy of 85% (Sharma et al., 2021). However, studies based on large clinical datasets are still limited, while few studies have systematically investigated the effects of individual variables such as the presence of sleep disorders, age, and gender in subjects.

Previous studies with sleep EEG analysis have revealed that even healthy individuals without sleep disorders may have different characteristics depending on their age. It has been found that not only does the time consumed for slow-wave sleep (SWS) decrease with aging but also the power of the activity itself during SWS is reduced (Landolt et al., 1996; Campos-Beltrán and Marshall, 2021). In addition, the sleep spindle, which is the key feature of the N2 stage (Werth et al., 1997; Fogel and Smith, 2011), has been found to decrease in its amplitude, density, and length with aging (Campos-Beltrán and Marshall, 2021). Therefore, in building a model that automatically evaluates sleep EEG, it is necessary to consider the age of subjects in the training dataset and to systematically analyze the effect thereof.

In this study, we first built a machine learning model that automatically performs sleep stage classification using the PSD features of sleep EEG and three different algorithms: SVM, kNN, and multilayer perceptron (MLP). In particular, we analyzed the age-related effects by constructing a general model trained on the data of all subjects regardless of their age and age-specific models (younger and older group models) and compared their performance. We then conducted OSA screening based on a model trained on EEG features

for each sleep stage and evaluated its feasibility. Therefore, we provide a comprehensive computational framework that automatically scores sleep stages and further determines the risk of OSA.

2. Methods

2.1. Data acquisition and preprocessing

Data from a total of 139 subjects from standard PSG conducted at Ewha Womans University Mokdong Hospital were employed in this study (Table 1). The age of the subjects ranged from 18 to 65 years. To investigate age-related effects, we divided the subjects into two groups: a younger group aged 18–45 years, including young and early middle-aged adults, and an older group aged 46–65 years, including late middle-aged adults (Medley, 1980). The number of subjects in the younger and older groups were 69 and 70, respectively. The dataset included 42 healthy subjects and 97 patients with OSA. The patients with OSA were further subdivided into mild to moderate (mtom) and severe groups according to clinical indices evaluating the severity (Gul et al., 2018), and the values in each group for Respiratory Disturbance Index (RDI) and blood oxygen saturation (SpO₂), which are key clinical indices, are presented in Table 1. To examine only the effects of OSA, all subjects with other medical histories such as stroke, neurological disorders, alcoholism, cancer, hypertension, and thyroid problems were excluded. More detailed demographic characteristics for the 139 subjects are provided in Table 1.

Data from all subjects were collected with Twin PSG Clinical Software (Glass Technologies, Warwick, RI, United States) (Choi et al.,

TABLE 1 Demographic characteristics of the subjects involved in this study.

	Younger group (aged 18–45 years)	Older group (aged 46–65 years)
Numbers of subjects	69	70
Age (mean ± SD)	32.91 ± 8.9	52.47 ± 7.22
OSA diagnosis (N, RDI, SpO ₂)		
Healthy	22 (RDI = 4.35 ± 15.33, SpO ₂ = 89.40 ± 8.07)	20 (RDI = 5.97 ± 10.03, SpO ₂ = 88.4 ± 6.41)
Mild to moderate OSA	25 (RDI = 15.90 ± 9.07, SpO ₂ = 82.92 ± 4.39)	24 (RDI = 12.46 ± 11.12, SpO ₂ = 83.70 ± 5.84)
Severe OSA	22 (RDI = 53.51 ± 26.61, SpO ₂ = 73.40 ± 10.95)	26 (RDI = 51.05 ± 20.15, SpO ₂ = 74.92 ± 7.90)
Gender (N)		
Female	8	24
Male	61	46
Body mass index (mean ± SD)	27.12 ± 5.23	26.39 ± 3.95
Smoking status (N)		
Never	42	42
Ex-smoker	7	15
Current	20	13
Number of epochs for each sleep stage		
(total, mean ± SD)		
Wake	8,027 (116.33 ± 103.16), 14.24%	8,218 (117.40 ± 72.13), 14.63%
N1	12,234 (117.30 ± 118.89), 21.71%	13,365 (190.93 ± 117.64), 20.23%
N2	25,799 (373.90 ± 122.47), 45.78%	25,167 (359.52 ± 114.79), 44.81%
N3	2,502 (36.26 ± 47.24), 4.44%	1,181 (16.87 ± 30.70), 2.10%
REM	7,796 (112.90 ± 50.90), 13.83%	8,235 (117.64 ± 51.73), 14.66%

2021). Full night sleeps of approximately 5–7 h were recorded for each subject. Sleep EEG was recorded using six electrodes (F3, F4, C3, C4, O1, and O2) following the international 10–20 system. A single ground electrode was attached on the forehead, and two linked ear electrodes were used as references. Impedances of electrodes were kept under 10 k Ω and the sampling rate was 200 Hz. A more detailed description of PSG data acquisition can be found in our previous paper (Choi et al., 2021). The sleep EEG was segmented into 30-s epochs, and each epoch was scored as one of five sleep stages (wake, N1, N2, N3, REM), as inspected by experts following the AASM manual (version 2.6; Berry et al., 2020).

The artifacts of EEG tend to contribute to the features of distinct sleep stages (Brunner et al., 1996); therefore, preprocessing of the sleep EEG was performed simply with bandpass filtering of 0.5–50 Hz. Out of data from 139 subjects, data from 111 subjects were used for the training set, and the remaining data (from 28 subjects) were used for the test set. The training and test sets were evenly distributed by age and OSA diagnosis.

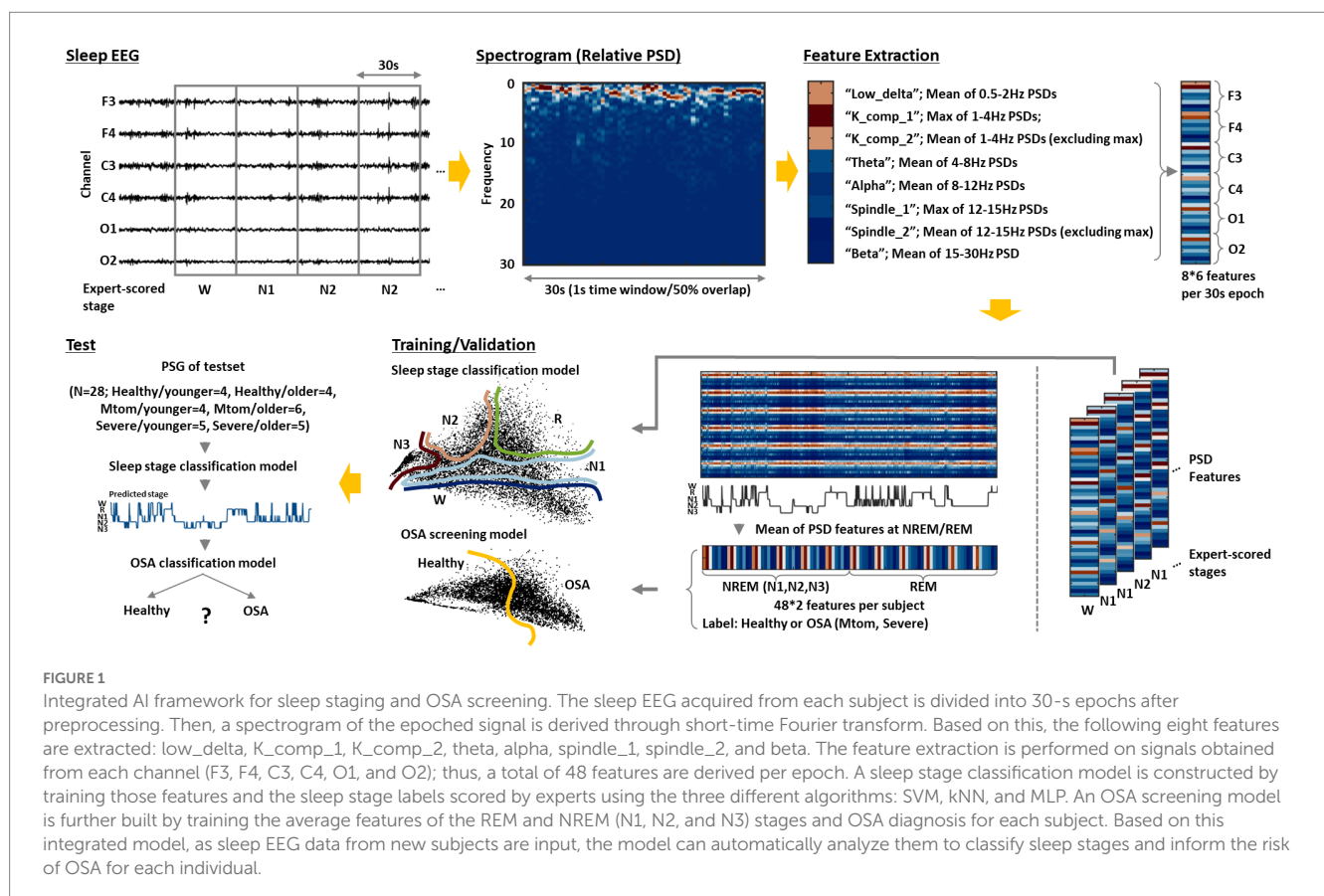
2.2. Integrated AI framework for sleep staging and OSA screening

We propose an integrated AI framework that automatically classifies sleep stages by analyzing the acquired sleep EEG and further screens OSA based on those results. The overall pipeline on this is depicted in Figure 1. The preprocessed sleep EEG was divided into 30-s epochs in the same manner as experts review, and eight main

features were extracted from the spectrogram obtained *via* time-frequency analysis of the signal at each epoch. The features were extracted from the signals for each channel. A more detailed description for feature extraction is provided in the following section. A sleep stage classification model was built by training the features for each epoch and the labels (sleep stage) scored by experts based on three different algorithms: SVM, kNN, and MLP. An OSA screening model was built by training the eight features in the REM and NREM (N1, N2, N3) stages of the healthy and OSA groups using the corresponding algorithms. In particular, to examine the age-related effect, a general model trained on the data of all subjects regardless of their age and age-specific models trained on only data from each age group were separately constructed, and their performance was compared. Each model was verified through five-fold cross-validation repeated 10 times, and the performance of the final model was further evaluated using the test set that was uniformly extracted according to the age and OSA diagnosis.

2.3. Feature extraction

For model training for sleep stage classification, we employed frequency-domain features that can directly quantify specific patterns crucial for discriminating sleep stages with relatively simple computations (Aboalayon et al., 2016; Malafeev et al., 2018). According to the AASM manual, the epoch is labeled ‘wake’ if 50% or more of the signal has an alpha rhythm. Stage N1 is scored when the alpha rhythm is attenuated and replaced with lower amplitude, mainly



theta power. Stage N2 is similar to N1, except that the epoch has unique features called a K complex and sleep spindle (Krauss et al., 2018; Ioannides et al., 2019). The K complex refers to a large abrupt activity with a delta frequency component, and the sleep spindle refers to a brief burst of the sigma band. They occur approximately every two epochs and have a duration of 0.5–1.5 s (Huang et al., 2014). Finally, stage N3 is scored when the slow waves of delta band continue (particularly low delta components of 0–2 Hz (Huang et al., 2014)).

To consider the abovementioned characteristics, we first performed a short-time Fourier transform (STFT) using MATLAB (version R2019b) to derive the spectrogram of the signal at each epoch. The STFT was calculated based on time windows of 1 sec with 50% overlap, resulting in 59 PSD vectors in one epoch (30 s). PSD vectors derived from each time window were normalized by their sum. The frequency band was divided into low delta (0.5–2 Hz), delta (1–4 Hz), theta (4–8 Hz), alpha (8–12 Hz), sigma (12–15 Hz), and beta (15–30 Hz) bands, and eight features to be used for model training were extracted from these results. In particular, to detect the K complex and sleep spindle, maximum values of 59 relative PSDs were extracted in the delta and sigma bands, respectively, and the mean values of the remaining 58 values were also extracted (Huang et al., 2014). For other frequency bands, an average value of 59 relative PSDs was extracted (Figure 1). Consequently, a total of 48 features were derived per epoch because eight features were extracted from the signals of each channel (F3, F4, C3, C4, O1, and O2).

Referring to a previous paper that demonstrated significant differences in sleep EEG characteristics in REM and NREM stages between healthy and OSA groups (Kang et al., 2021), we derived the average features corresponding to the REM and NREM (including N1, N2, and N3) stages based on the extracted features for each epoch, and used those features for OSA screening model training.

2.4. Model training

Both the sleep stage classification and the OSA screening models were trained with a supervised machine learning approach that learns the extracted EEG features and labels assigned by experts (sleep stages or OSA diagnosis). For the model training, three different algorithms were employed including SVM, kNN, and MLP.

Regarding the kernel for the SVM, the radial basis function (RBF) was used (Huang et al., 2014). The kernel trick is remapping to a different plane or dimension to obtain a decision boundary, wherein the SVM operates with two hyperparameters called C and gamma. A grid search was performed to set the parameter values, within a set C [0.1, 1, 10, 100] and gamma [0.001, 0.01, 0.1, 1, 10], and the best parameters calculated per the experiments were applied. With respect to the hyperparameters of kNN, the number of neighbors ($n_{\text{neighbors}}$), weights between the neighbors, and metric used for distance calculation were adjusted, and the best parameter values acquired from the grid search, within a set $n_{\text{neighbors}}$ [1, 3, 5, 7, 9, 11, 13, 15, 17, 19], weights ['uniform', 'distance'], and metric ['euclidean distance', 'manhattan distance'], were used. For the MLP, the number of hidden layers and the number of nodes for each hidden layer that determine the neural network structure were used as hyperparameters. The optimal combination was derived and used through the grid search, within a set the number of hidden layers [1, 2] and the number of nodes for each hidden layer [16, 32, 64]. Weight optimization at

each node was performed based on Adam, which is a stochastic gradient-based optimization algorithm (Kingma and Ba, 2014), and a rectified linear unit activation function, and the maximum number of iterations was set to 200.

2.5. Model validation and performance evaluation

Five-fold cross-validation was repeated 10 times on the built models, and the performance between the models was statistically analyzed based on the 50 accuracy values derived from this assessment. Comparisons between the built models or groups were performed through one-way ANOVA, which is based on the F-statistic (Verma, 2013). ANOVA results are presented as p -values and F -values with degrees of freedom (between and within groups). The significance level was defined as $p < 0.01$, and post-hoc analyses were performed for significant results using independent-samples t -test with Bonferroni correction.

The performance of the final models was further assessed by employing the test set evenly extracted depending on age and OSA diagnosis.

3. Results

3.1. Sleep stage classification

3.1.1. Performance comparison between the general model and age-specific models

The sleep stage classification models were built based on three different learning algorithms (SVM, kNN, and MLP), and each model was evaluated *via* five-fold cross-validation repeated 10 times, in which each iteration was performed based on a newly shuffled dataset (Figure 2A). To investigate the age-related effects, two types of models were built with each learning algorithm, and their performance was assessed: a general model trained with data from all subjects regardless of age and age-specific models trained only with data from younger or older subjects. The age-specific models were validated using subjects' data in their own age-group (Y-Y, O-O) as well as data from the other age-group (Y-O, O-Y).

The classification accuracy for each sleep stage (wake, N1, N2, N3, REM) and the overall accuracy across all stages of the abovementioned models are presented in Figure 2B. In the SVM-based models, statistical analyses demonstrated significant differences among the accuracies of the general model and age-specific models [in all stages: $F(2,147) = 15.27$, $p < 0.001$; wake: $F(2,147) = 7.35$, $p = 0.239$; N1: $F(2,147) = 8.88$, $p < 0.001$; N2: $F(2,147) = 35.99$, $p < 0.001$; N3: $F(2,147) = 141.92$, $p < 0.001$; REM: $F(2,147) = 14.64$, $p = 0.002$], and group pairs with significant differences in each post-hoc analysis are indicated by gray lines in Figure 2B (top row). The overall performance of the younger group model showed no noticeable difference with that of the general model, whereas the older group model revealed significantly poorer performance than the general model (mean \pm SD of $f1$ -scores in all stages: general model = 0.69 ± 0.02 , Y-Y = 0.69 ± 0.03 , O-O = 0.67 ± 0.02 , $p < 0.001$ for the general model $>$ O-O). The results for each sleep stage showed similar patterns to the overall outcome for all stages. The young group model demonstrated better (in N3: general

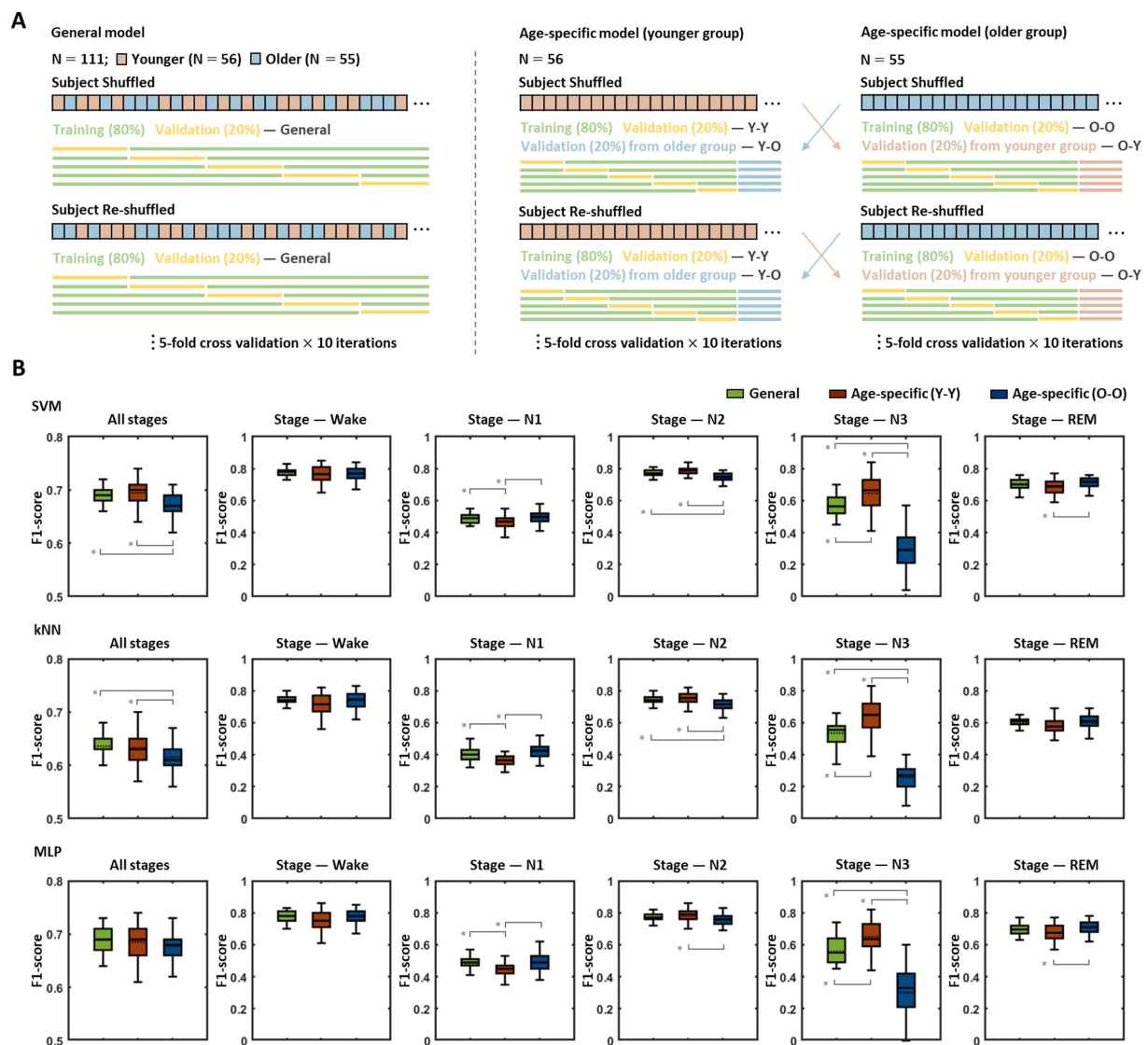


FIGURE 2

Performance comparison between the general model and age-specific models. **(A)** Model training and validation procedures. Each model was assessed by five-fold validation repeated 10 times, and every five-fold validation was conducted based on a newly shuffled dataset. The general model was trained and validated based on the data from all subjects, the age-specific models were trained with the data corresponding to each age group, and the validation was performed using data belonging to their own age group (Y-Y, O-O) as well as data from the other age group (Y-O, O-Y). **(B)** Validation results for the general model and age-specific models. The figures show the results obtained by the validation procedures of each model for each sleep stage. Pairs with statistically significant differences in classification accuracy are indicated by solid gray lines. An asterisk represents a significant difference ($* p < 0.01$).

model = 0.56 ± 0.08 , Y-Y = 0.65 ± 0.11 , $p < 0.001$ for the general model < Y-Y) or similar performance (in wake: general model = 0.78 ± 0.03 , Y-Y = 0.77 ± 0.05 ; N2: general model = 0.77 ± 0.02 , Y-Y = 0.79 ± 0.03 ; REM: general model = 0.70 ± 0.03 , Y-Y = 0.68 ± 0.05) compared with the general model, except for the N1 stage (general model = 0.49 ± 0.03 , Y-Y = 0.47 ± 0.04 , $p = 0.001$ for the general model > Y-Y). On the other hand, the older group model exhibited significantly lower accuracy in N2 and N3 stages than the general model (in N2: O-O = 0.74 ± 0.03 , $p < 0.001$ for the general model > O-O; N3: O-O = 0.30 ± 0.13 , $p < 0.001$ for the general model > O-O). There were no statistically significant differences in the other stages (in wake: O-O = 0.77 ± 0.04 ; N1: O-O = 0.49 ± 0.04 ; REM: O-O = 0.71 ± 0.04). In addition, with respect to age-specific models, the accuracy when using the validation set

from their own age group was relatively higher than that obtained when using the validation set from the other age group (in all stages: Y-O = 0.68 ± 0.02 , O-Y = 0.66 ± 0.03 , $p = 0.002$ for Y-Y > Y-O, $p = 0.029$ for O-O > O-Y; [Supplementary Figure S1](#)).

The patterns of performance differences among the general model and age-specific models identified in the kNN-based models were fairly similar to those identified in the SVM-based models ([Figure 2B](#), middle row). In particular, in terms of overall accuracy for all sleep stages, a statistically significant difference was derived among the models [in all stages: $F(2,147) = 9.11$, $p < 0.001$], and while the younger group model showed no significant difference from the general model, the older group model demonstrated significantly lower accuracy than the general model (in all stages: general model = 0.64 ± 0.02 ,

Y-Y = 0.63 ± 0.03 , O-O = 0.62 ± 0.03 , $p < 0.001$ for the general model > O-O).

In contrast to the findings in the SVM- and kNN-based models, in MLP-based models, no significant performance differences were identified among the models by age group [in all stages: $F(2,147) = 2.30$, $p = 0.104$; Figure 2B, bottom row]. Compared with the general model, both age-specific models demonstrated no significant difference in the overall accuracy across all stages (in all stages: general model = 0.69 ± 0.02 , Y-Y = 0.69 ± 0.03 , O-O = 0.68 ± 0.02), even though the younger group model and older group model revealed lower accuracy than the general model for the N1 and N3 stages, respectively (in N1: general model = 0.48 ± 0.04 , Y-Y = 0.45 ± 0.05 , $p < 0.001$ for general model > Y-Y; N3: general model = 0.56 ± 0.10 , O-O = 0.30 ± 0.15 , $p < 0.001$ for general model > O-O).

The classification accuracy for sleep stage also demonstrated a significant difference depending on the learning algorithm (in all stages using the general model: $F(2,147) = 130.57$, $p < 0.001$). The SVM- and MLP-based model showed similar levels of accuracy, whereas the kNN-based model yielded significantly lower accuracy than those two models (SVM-based model = 0.69 ± 0.02 , kNN-based model = 0.64 ± 0.02 , MLP-based model = 0.69 ± 0.02 , $p < 0.001$ for the SVM-based model > kNN-based model, $p < 0.001$ for the MLP-based model > kNN-based model). This difference in performance among the models using the different algorithms was consistently identified not only across all stages but also for each sleep stage except for the N3 stage [in wake: $F(2,147) = 18.17$, $p < 0.001$; N1: $F(2,147) = 89.71$, $p < 0.001$; N2: $F(2,147) = 20.80$, $p < 0.001$; N3: $F(2,147) = 1.37$, $p = 0.257$; R: $F(2,147) = 160.23$, $p < 0.001$].

3.1.2. Sleep stage prediction results using the general model

The results of the previous section indicated that the age-specific model may have poorer performance than the general model depending on the age groups. Therefore, we selected the age-integrated general model as a representative model for sleep stage classification and employed it for subsequent work. With respect to the learning algorithm, all three different algorithms, SVM, kNN, and MLP, were used independently.

Figure 3A depicts the hypnograms for a single subject (healthy older adult) among a test set of 28 subjects. The upper red graph indicates the expert scored sleep stages, and the lower blue graph shows the automatically classified stages by the SVM-based model. Prediction via the model demonstrated an overall similar hypnogram for ~7 h of total sleep and yielded an accuracy (f1-score) of 0.82 in this particular subject. The test results for a total of 28 subjects in each model (age-integrated general model) trained based on the three different learning algorithms are presented in Figures 3B,C. The SVM-, kNN-, MLP-based models revealed average accuracies of 0.72, 0.67, and 0.73, respectively, for all stages. For each stage, they showed high accuracies for the wake and N2 stages but relatively low accuracies for the N1 and N3 stages. Consistent with the model validation results in the previous section, the kNN-based model demonstrated relatively poorer performance than the SVM- and MLP-based models in terms of accuracy at each stage and overall accuracy across all stages; however, in the current test results, the difference in accuracy among the models did not reach statistical significance. The confusion matrices (Figure 3C) visualize the

performance of each model at a glance and inform the ratios of the predicted sleep stages to the actual stages (expert-scored sleep stages).

3.2. Obstructive sleep apnea screening

3.2.1. Differences in EEG features according to the presence and severity of OSA

Patients with OSA diagnosed based on multiple clinical indices, including respiration-related indices, demonstrated marked differences in the characteristics of sleep EEG as well as in corresponding indices compared with healthy subjects. Such differences in EEG features according to the presence or severity of OSA also differed by age group. Figure 4 shows eight EEG features of healthy subjects, patients with mtom OSA, and patients with severe OSA for the NREM (including N1, N2, and N3) and REM sleep stages in the younger and older groups.

Statistical analysis in the younger group demonstrated significant differences in the K complex, beta-band, and spindle features according to OSA severity [in NREM stages: K_comp_1: $F(2,53) = 7.61$, $p = 0.001$; K_comp_2: $F(2,53) = 7.67$, $p = 0.001$; Beta: $F(2,53) = 6.10$, $p = 0.004$; in REM stages: K_comp_1: $F(2,53) = 8.90$, $p < 0.001$; K_comp_2: $F(2,53) = 7.31$, $p = 0.002$; Spindle_1: $F(2,53) = 5.10$, $p = 0.009$; Spindle_2: $F(2,53) = 6.21$, $p = 0.004$; Beta: $F(2,53) = 7.01$, $p = 0.002$]. In particular, the K complex features (K_comp_1, K_comp_2) exhibited a tendency to decrease toward severe OSA in both the NREM and REM sleep stages, and statistically significant pairs are indicated by gray lines. In contrast, although there was no significant difference between the healthy and mtom OSA groups, the beta-band features showed a tendency to increase toward severe OSA in both sleep stages. In the REM stages, significant differences were elicited in the spindle features (spindle_1 and spindle_2) in addition to the K complex and beta-band features; there was no significant difference between the healthy and mtom OSA groups, but the severe OSA group showed significantly higher values than those two groups.

In contrast to the younger group, differences in EEG characteristics depending on OSA severity were not clearly observed in the older group. In particular, in the NREM stages, K complex and beta-band features represented similar tendencies to those identified in the younger group [only the beta-band features reached statistical significance; $F(2,53) = 7.57$, $p = 0.001$], but no between-group differences (according to OSA severity) were elicited in the REM stages.

3.2.2. Performance analysis of the model

The OSA screening model was built by training the EEG features in the REM/NREM stages (expert-scored stages) and OSA diagnostic results for each subject based on the three learning algorithms. As in the previous section (section 3.1), to examine the age-related effect, the general model and age-specific models were independently generated, and their performance was compared. The models were further evaluated by comparison with models trained on SpO₂, a respiration-related measure, along with EEG features, to investigate the feasibility of using EEG features.

The average accuracies of the OSA classification (healthy vs. mtom OSA vs. severe OSA) in the general models were 0.46, 0.47, and 0.50 for the SVM, kNN, and MLP algorithms, respectively, when evaluated by performing 10 iterations of five-fold cross-validation; these were

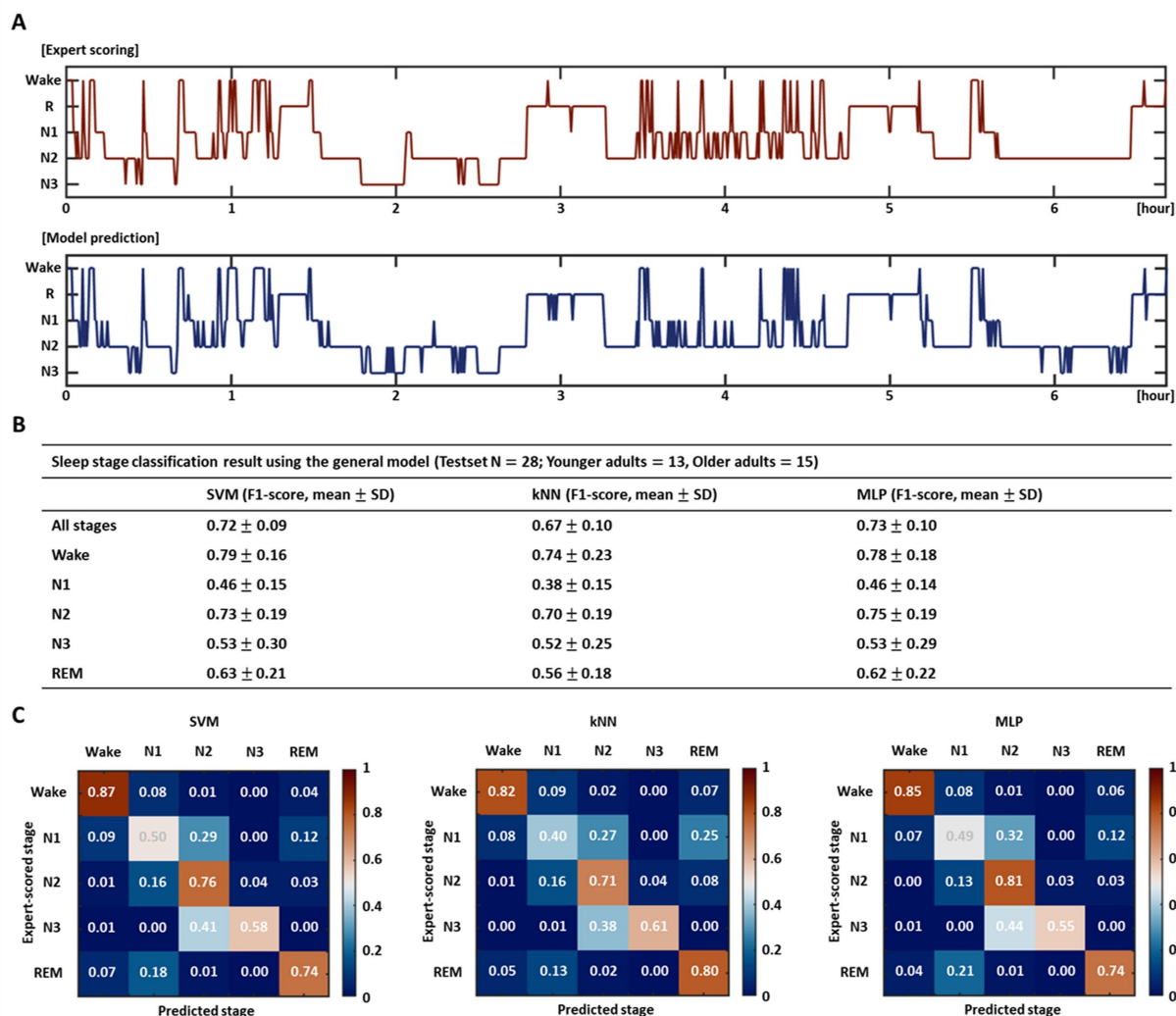


FIGURE 3 Performance evaluation of sleep stage classification model. **(A)** Test result of a particular subject using the SVM-based general model. The upper red hypnogram represents the expert scored sleep stages, and the lower blue hypnogram indicates the automatically classified stages by the model. **(B)** Sleep stage classification accuracy using the entire test set for each model trained by three different learning algorithms and **(C)** their confusion matrices. The table exhibits the overall accuracy for all stages and the accuracy for each stage. The confusion matrices show the ratios of the predicted stages to the expert-scored stages.

far below the average accuracies of the models trained on the EEG features and SpO₂ (0.64, 0.67, and 0.64 for the SVM, kNN, and MLP algorithms, respectively) (Table 2). However, in the case of the models trained excluding the features of the mtom OSA group, which had intermediate characteristics between the healthy and severe OSA groups, the accuracies considerably increased to 0.74, 0.69, and 0.74 for the SVM, kNN, and MLP algorithms, respectively, even though they did not reach those of the models based on both EEG features and SpO₂ (0.86, 0.87, and 0.89 for the SVM, kNN, and MLP algorithms, respectively). These patterns were consistently observed in age-specific models as well as the general models.

With respect to the age-related effect, no statistically significant differences were elicited in the performance of the general and age-specific models across all the three learning algorithms. However, in terms of the OSA classification, the age-specific models tended to show slightly lower accuracies than the general model, and the older group model had the

lowest accuracy (Table 2). In the healthy vs. severe OSA groups, the younger group model derived a similar accuracy to the general model, but the older group model yielded a relatively lower accuracy than those two models; these characteristics were more pronounced in models trained using the SVM or kNN algorithms.

Finally, regarding the effect of learning algorithms, although MLP-based models demonstrated slightly better performance overall than SVM- and kNN-based models, there was no statistically significant difference.

3.2.3. OSA prediction results using the integrated model combining the sleep stage classification model and OSA screening model

Referring to the results of the analysis in the previous section, the age-integrated general model trained excluding the features of the mtom OSA group was defined as an OSA screening model

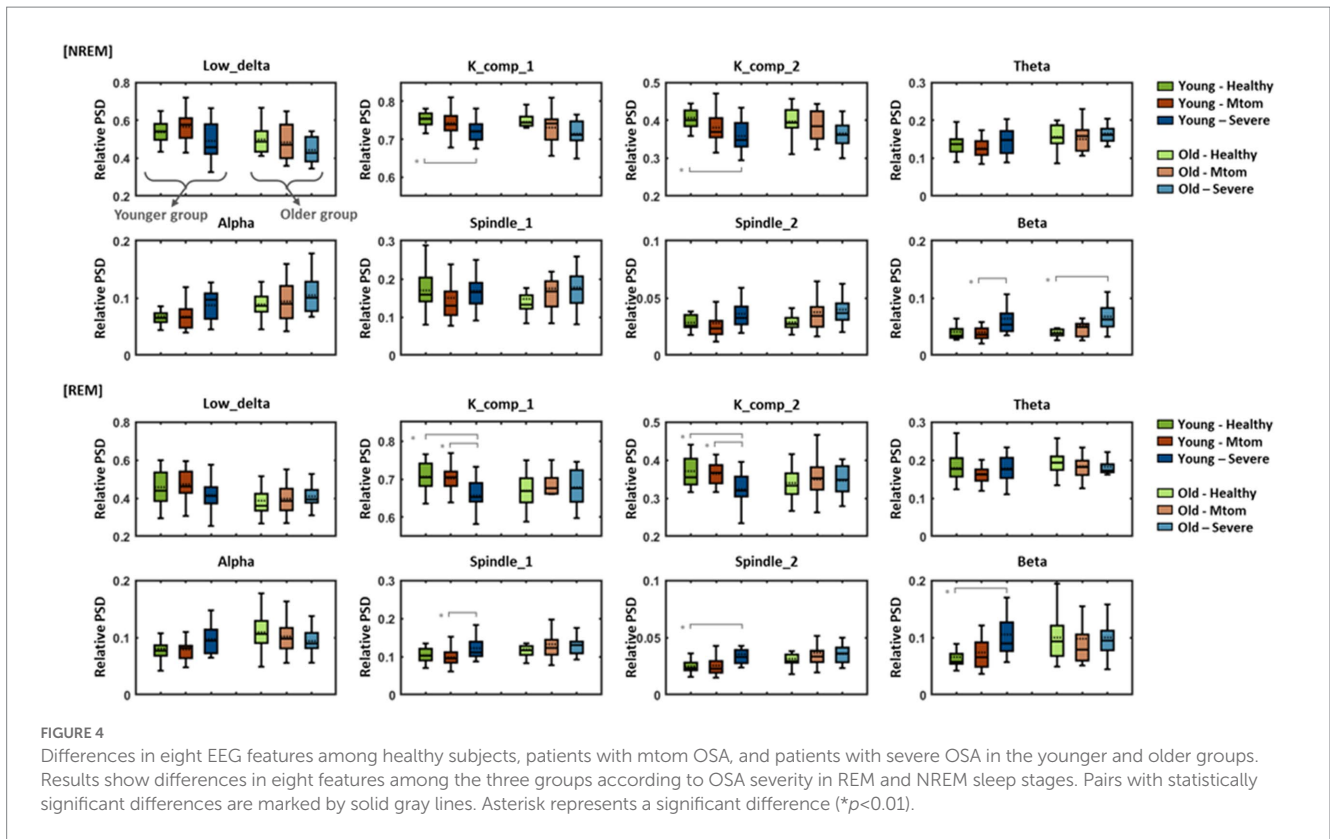


TABLE 2 Validation of OSA screening model.

Validation results of the OSA screening model (10 iterated five-fold cross validation, N = 111; Healthy = 34, Mtom OSA = 39, Severe OSA = 38)

	General		Age-specific (Y-Y)		Age-specific (O-O)	
	Model trained with REM/NREM EEG features	Model trained with REM/NREM EEG features + SpO ₂ index	Model trained with REM/NREM EEG features	Model trained with REM/NREM EEG features + SpO ₂ index	Model trained with REM/NREM EEG features	Model trained with REM/NREM EEG features + SpO ₂ index
SVM						
Healthy vs. Mtom OSA vs. Severe OSA	0.46 ± 0.10	0.64 ± 0.09	0.44 ± 0.13	0.67 ± 0.15	0.43 ± 0.13	0.57 ± 0.16
Healthy vs. Severe OSA	0.74 ± 0.12	0.86 ± 0.08	0.74 ± 0.17	0.88 ± 0.08	0.70 ± 0.14	0.83 ± 0.11
kNN						
Healthy vs. Mtom OSA vs. Severe OSA	0.47 ± 0.11	0.67 ± 0.10	0.42 ± 0.16	0.67 ± 0.13	0.42 ± 0.15	0.55 ± 0.15
Healthy vs. Severe OSA	0.69 ± 0.11	0.87 ± 0.07	0.70 ± 0.19	0.88 ± 0.11	0.62 ± 0.15	0.84 ± 0.12
MLP						
Healthy vs. Mtom OSA vs. Severe OSA	0.50 ± 0.11	0.64 ± 0.10	0.48 ± 0.14	0.66 ± 0.15	0.45 ± 0.14	0.55 ± 0.15
Healthy vs. Severe OSA	0.74 ± 0.12	0.89 ± 0.08	0.71 ± 0.16	0.90 ± 0.11	0.72 ± 0.15	0.85 ± 0.12

(healthy vs. OSA), and its performance was evaluated by applying the test set. The test was conducted in a manner that first automatically classified the sleep stages using the model built in the previous section 3.1 and then applied the results to the OSA screening model. The outcomes revealed that the accuracy varied depending on the learning algorithms used in training the sleep

stage classification model and OSA screening model (Table 3). In particular, the highest accuracy of 0.73 was derived when the MLP algorithm was applied to both models for training, which indicates that patients with OSA could be screened with the corresponding accuracy level only with sleep EEG characteristics without any respiration-related measures.

TABLE 3 OSA screening results.

OSA screening result using the general model (Testset $N = 28$; Younger adults = 13, Older adults = 15)									
Sleep stage classification model	SVM	SVM	SVM	kNN	kNN	kNN	MLP	MLP	MLP
OSA screening model	SVM	kNN	MLP	SVM	kNN	MLP	SVM	kNN	MLP
	0.65	0.45	0.71	0.58	0.52	0.71	0.65	0.53	0.73

4. Discussion

We have proposed an integrated computational framework that can automatically analyze sleep EEG data obtained from each subject, classify sleep stages using machine learning techniques, and determine the risk of OSA based on these findings. The current outcomes have demonstrated the feasibility of AI technologies that can play a beneficial role in clinical applications by enabling automated and systematic analyses.

With respect to automatic sleep stage classification, we constructed a general model that trained on data from all subjects regardless of age and age-specific models that trained only on data from each age group and compared their performance. As a result, the younger group model showed similar accuracy to the general model and even higher accuracies in some stages, although they did not reach statistical significance. In contrast, the older group model exhibited significantly lower accuracies than the general model. In the case of the age-specific models, when the data of subjects belonging to their own age group were used for validation, the accuracy was generally higher than when data belonging to the other age group were used. All these results suggest that there are some differences in sleep EEG characteristics between age groups, as evidenced by the findings of previous studies (Landolt et al., 1996; Campos-Beltrán and Marshall, 2021). In particular, the results imply that the sleep EEG data of younger populations have homogeneous characteristics, given that the younger group model showed similar or higher accuracies despite the smaller training sets than the general model. On the other hand, the performance degradation of the older group model may be interpreted as individual variability increases as aging progresses; thus, the older population has more heterogeneous characteristics.

Previous studies have demonstrated that the older group has larger inter-individual variability than the younger group in terms of not only macro-level sleep architecture (including total sleep time, sleep efficiency, and the ratio of time spent in each sleep stage) but also micro-level architecture that can be identified from the EEG, such as spindle density and REM density (Peters et al., 2014; Mander et al., 2017). Moreover, such characteristics were also observed in the current data set. As a result of examining inter-individual variability based on the Pearson correlation coefficient using the PSD features of each subject's sleep EEG, a significant difference was found between the two age groups; in general, higher correlation values were derived between individuals in the younger group than in the older group, indicating that the features in the younger group are more homogeneous (Supplementary Figure S2). These group differences were especially pronounced in the wake, REM, and N1 stages. In contrast, in the N3 stage, the older group showed higher correlation values; however, it should be considered that the number of samples for the N3 stage in the older group was significantly smaller than that of the younger group due to the nature of reduced deeper NREM sleep in the older population. The results emphasize that attention should

be paid to the bias of individual factors, such as age bias, when training a model using sleep EEG data. Furthermore, future studies are needed to systematically investigate aging-related effects on sleep, especially in terms of changes in brain network characteristics.

The sleep stage scoring accuracy also differed depending on the algorithms used to train the model. Based on the age-integrated general model, the test results for all stages yielded average accuracies of 72, 67, and 73%, when applying the SVM, kNN, and MLP algorithms, respectively, and the SVM- and MLP-based models derived relatively higher performance than the kNN-based model across all stages and for each stage, although not reaching statistical significance. The classification accuracy of sleep stages revealed differences for each stage. While the accuracies for the wake and N2 stages were fairly high, the accuracies for the N1 and N3 stages were relatively low, and these characteristics were consistently observed in the models trained with three different algorithms.

Regarding the N3 stage, the accuracy was somewhat lower than that reported in previous studies (Huang et al., 2014; Zhu et al., 2014; Acharya et al., 2015), and it was often mispredicted as the N2 stage. This outcome may be due to the characteristics of the data set used in this study. To investigate the effects of age and OSA, in the current study, we included more data from patients with OSA as well as older adults compared to the existing studies. In other words, the proportion of older adults and patients with OSA in the entire data set is quite high. The N3 stage represents deep sleep that constitutes approximately 10–20% of total sleep time in healthy people (Kryger et al., 2017), which naturally decreases with aging or as sleep quality deteriorates due to sleep disorders, such as OSA. Therefore, the low accuracy of the model for the N3 stage is likely to be induced by insufficient learning of the features of the corresponding stage, along with increased individual variability with aging, depending on the demographic characteristics of the data set; The model often misclassified N3 as the N2 stage, which was relatively well-trained. This can be improved by acquiring more data sets and training them further.

The low accuracy for the N1 stage and especially the confusion with the N2 stage have been frequently observed in the existing machine learning studies for automatic sleep stage classification (Panossian and Avidan, 2009; Hsu et al., 2013; Supratak et al., 2017; Sors et al., 2018). In fact, both stages belong to shallow sleep and share similar characteristics. According to the AASM manual, the scoring for the N2 stage is based on the occurrence of a K-complex or sleep spindle. However, if the N2 stage has preceded beforehand, the following epoch is also scored as N2 in the absence of arousal or interruption, even if those two features are not observed. In other words, the scoring is performed not only by the features of the sleep EEG but also by the pre-post relationship of the sleep stages. The learning algorithms we applied here, which classify sleep stages by training the PSD features of the sleep EEG for each epoch, have a limitation in that they cannot reflect the scoring considering such a pre-post relationship, thereby exhibiting relatively low performance for the N1 stage. To overcome this issue, it may be an alternative way to use

an algorithm that can learn the pre-post data features and use them for stage scoring, such as bidirectional long short-term memory (Zhang et al., 2019; Kuo and Chen, 2020), even at an increased computational cost.

Table 4 presents a summary of the characteristics of several recently published notable EEG-based sleep stage classification models. As expected, models trained on data from healthy subjects with a small age variance demonstrated relatively high overall performance across different learning algorithms compared to other models (Ghasemzadeh et al., 2019). In particular, the model built with data from young, healthy subjects demonstrated the highest accuracy. Direct comparison may be unreasonable given that models can operate at different performance levels depending on the data set, even if the same training approach is applied (Fan et al., 2021); however, the current model exhibited accuracy similar to existing models built on data sets including subjects with sleep disorders and middle-aged adults (75% in Tzimourta et al., 2018, 72% in Tripathy et al., 2020, and 75% with the original unbalanced training approach in Sharma et al., 2021). Recent studies have applied additional processes during model training to improve the classification accuracy of the model. Sharma and colleagues performed unbiased training by equalizing the number of epochs used in each sleep stage learning (i.e., generating a balanced dataset) using over-sampling and under-sampling techniques to resolve the learning imbalance for each sleep stage, which elicited markedly improved performance compared to the training model on the original unbalanced dataset (85% in the balanced dataset and 75% in the original unbalanced dataset; Sharma et al., 2021). Hussain and colleagues used clean preprocessed EEG data by removing signal artifacts (including ocular, muscle, and motion artifacts) via independent component analysis and added a feature selection procedure based on further statistical analyses, thereby increasing the accuracy of the model (84–89%; Hussain et al., 2022). The current study aimed to create an EEG-based integrated model for sleep stage scoring and OSA screening and evaluate its feasibility under the condition of minimizing the amount of computation and complexity, considering compatibility with wearable devices and mobile applications, thus, we did not apply additional procedures to improve the model performance. It is expected that models with enhanced performance can be built by applying appropriate additional processes considering the trade-off with computational power, if necessary.

Characteristics of sleep EEG at each NREM and REM stage demonstrated significant differences between healthy subjects and patients with OSA, consistent with previous findings (Kang et al., 2021). Those differences were more pronounced in the younger group than in the older group. The younger group demonstrated significant differences in K complex and beta band features across NREM and REM stages according to the presence or severity of OSA (healthy vs. mtom OSA vs. severe OSA groups). The group showed additional significant differences in spindle features in the REM stage. The older group exhibited relatively similar patterns to the younger group in the NREM stage, but a significant difference was derived only for the beta band feature, and no noticeable differences (among the three groups) were found in the REM stage. Accordingly, the OSA classification model trained on EEG features in NREM and REM stages represented generally lower accuracy in the older age-specific model than the younger age-specific or general model. Concerning the learning algorithm used for model training, the MLP-based model yielded higher performance than the SVM- and kNN-based models, but the difference was not statistically significant.

The MLP-based OSA screening model trained excluding the features of the mtom OSA group, which has intermediate features that are relatively indistinguishable compared to the healthy or severe OSA groups, exhibited 73% of performance when applying the scoring results derived from the automatic sleep stage classification model. These results are critically meaningful in that the model was able to distinguish patients with OSA (mtom and severe OSA) from healthy subjects, using only the characteristics of the sleep EEG without respiration-related measures. The results further suggest the need for systematic studies for OSA disease in terms of brain networks, including whether the altered brain network properties cause respiration-related problems during sleep or vice versa. Regarding OSA screening, another feature that needs to be carefully investigated in sleep EEG is arousal. In fact, one of the main characteristics of patients with OSA is frequent arousals from sleep, which are accompanied by a sudden increase in EEG frequency (Altevogt and Colten, 2006; Yue et al., 2009). In the dataset used in the current study, a significant difference was observed in the number of arousals during sleep among the healthy, mtom, and severe OSA groups. Furthermore, the performance of the OSA screening improved when the model was trained by adding the number of arousals in addition to NREM/REM EEG features (Supplementary Figure S3). Although the current study did not cover the contents related to the arousal index in-depth, given that the recent findings demonstrated that gamma power in EEG arousal differs according to the severity of the respiratory event and sleep stages (Pitkänen et al., 2021), future studies may need to closely examine such characteristics and incorporate them into building the models as needed.

Although the current results showed the feasibility that the OSA screening could be achieved to some extent with only EEG characteristics, the accuracy still did not reach that of the model trained by adding the respiratory-related index, SpO₂, as features. This is in line with the results of a recent paper demonstrating that a deep learning-based model trained on pulse oximetry measures is effective in scoring sleep stages and estimating AHI (Huttunen et al., 2022), and suggests that SpO₂ monitoring may be important in increasing efficacy for the screening of, at least, OSA, among several other sleep disorders. Given that SpO₂ could be obtained through a relatively simple setup, such as a finger pulse oximeter, current state-of-the-art wearable devices, which are being developed for individual sleep quality evaluation (Koushik et al., 2019; Liao et al., 2020) may be expanded into integrated systems that include detecting SpO₂ levels as well as the EEG. Such systems would be particularly useful in enabling individuals to analyze their sleep patterns and assess their risk of sleep disorders, even at home.

In the current study, an integrated model was constructed that performs sleep stage classification and OSA screening based on sleep EEG, and its feasibility was verified. However, there are still several limitations, and further studies are needed. To investigate age-related effects, we divided the subjects into two age groups, built a model for each group or the entire group, and elicited meaningful results by comparing their performance, but due to the limitations of the data set, data from older adults aged 65 years or older were not included. Therefore, it is required to expand the model based on sleep EEG obtained from older adults and to further verify the current outcomes. In addition, due to the nature of OSA, where the rate of diagnosis is 3.3 times higher in men than women (Bixler et al., 2001), there was a gender bias in the current data set (men: 107, women: 32). Moreover,

TABLE 4 Summary of existing sleep stage classification models based on conventional machine learning techniques.

Study	Dataset	Subjects N, age, gender	Data channels	Features	Algorithm	Class (the number of epochs for each class)	Accuracy
Tzamourta et al. (2018)	ISRUC-sleep dataset	100 subjects with sleep disorders, 51 ± 16 (mean ± SD), Female: 45 Male: 55	6 channels (F ₃ -A ₂ , C ₃ -A ₂ , O1-A ₂ , F ₄ -A ₁ , C4-A ₁ , O ₂ -A ₁)	Energy for each of 5 subbands ^a	Naive Bayes	Wake, N1, N2,	56%
					Decision Tree	N3, REM	66%
					kNN	(20,104, 11,104,	65%
					SVM	27,398, 17,325,	67%
					Random Forests	11,256)	75%
Ghasemzadeh et al. (2019)	ISRUC-sleep dataset	10 healthy subjects, 40 ± 10, Female: 1, Male: 9	1 channel (C ₃ -A ₂)	Entropy for each of 11 subbands ^b (non-normalized log energy)	SVM	Wake, N1, N2,	82%
					EBT	N3, REM (1702,	81%
					weighted kNN	1,123, 2,850,	81%
						1976, 1,238)	
	Sleep-EDF dataset	8 healthy subjects, 28 ± 5, Female: 4, Male: 4	2 channels (Pz-Oz, Fpz-Cz)	Entropy for each of 11 subbands ^b (non-normalized log energy)	SVM	Wake, S1, S2, S3,	93%
						S4, REM (8,055, 604, 3,621, 672, 627,1,609)	
Tripathy et al. (2020)	CAP sleep dataset	25 subjects including 6 healthy subjects and 19 subjects with sleep disorders, 48 ± 14 for female, 64 ± 18 for male, Female: 10, Male: 15	4 channels (P ₄ -O ₂ , C ₄ -A ₁ , F ₄ -C ₄ , C ₄ -P ₄)	Dispersion entropy, Bubble entropy for each of 5 subbands ^c	Hybrid classifier (based on class-specific residuals using sparse representation and distances from nearest neighbors)	Wake, S1, S2, S3,	72%
						S4, REM (2,613,	
						1,537, 4,955,	
						2,707, 2,601,	
						2,947)	
Sharma et al. (2021)	CAP sleep dataset	80 subjects including 6 healthy subjects and 74 subjects with sleep disorders, 48 ± 20, Female: 32, Male: 48	2 channels (F ₄ -C ₄ , C ₄ -A ₁)	Norm features (I_1, I_1, I_0) and their statistics from 6 subbands signals (obtained from wavelet decomposition)	Decision tree	Wake, S1, S2, S3,	With EBT, for unbalanced training: 75%, for balanced training: 85%
					Logistic regression	S4, REM for original unbalanced epochs: (15,841,	
					Naive Bayes	3,519, 28,628,	
					SVM	8,804, 10,188,	
					kNN	13,687) for balanced epochs: (13,407, 13,407,	
Hussain et al. (2022)	HMC sleep staging dataset	100 subjects with sleep disorders, 54 ± 15 Female: 66, Male: 88	3 channels (F ₄ , C ₄ , O ₂)	Mean power, median frequency, mean frequency, spectral edge, peak frequency for each of 5 subbands ^d , delta-alpha power ratio, delta-theta power ratio, slow-fast wave power ratio	C5.0 decision tree	Wake, S1, S2, S3,	87%
					MLP	S4, REM	89%
					CHAID		84%

^aSubbands, 0–4, 4–8, 8–13, 13–30, and 30–60 Hz; ^bsubbands, 0.4–4, 4–8, 8–10, 10–13, 13–18, 18–25, 25–30, 30–36, 36–41, 41–46, and 46–50 Hz; ^csubbands, 0–4, 4–8, 8–13, 13–30, and 30–75 Hz; ^dsubbands, 0.5–4, 4–8, 8–13, 13–30, and 30–44 Hz; ISRUC, Institute of Systems and Robotics, University of Coimbra; CAP, Cyclic Alternating Pattern; HMC, Haaglanden Medisch Centrum; kNN, k-Nearest Neighbors; SVM, Support Vector Machine; EBT, Ensemble Bagged Trees; MLP, MultiLayer Perceptron; CHAID, Chi-Squared Automatic Interaction Detector.

there was a difference in the gender ratio between the two age groups. Such a gender bias may have affected the current results, as previous studies have demonstrated significant differences in sleep EEG characteristics depending on gender groups (e.g., higher PSDs in women, especially in delta, theta, low alpha, and high spindle frequency bands; Carrier et al., 2001; Simor et al., 2013). Thus, it is necessary to systematically analyze the effects of gender by using data sets without gender bias or by comparing models built according to gender.

As mentioned earlier, we here used relatively less complex and low-computational training approaches to determine the feasibility of the integrated model. The accuracy of the model could be further improved by removing signal artifacts in the sleep EEG, selecting only clean epochs and applying them to the model training, or employing a more complex architecture, such as recurrent neural networks that can reflect time-varying dynamic features. In addition, regarding the features used for model training, frequency-domain features, which can directly quantify important patterns in sleep stage scoring with simple computations, were used in the current study. However, given that previous studies used various types of features including time-domain and nonlinear features in addition to frequency-domain features, and demonstrated quite good model performance (Aboalayon et al., 2016), further studies are needed to investigate the effects of features by performing comparative analysis on diverse features. Applying dimensionality reduction or feature selection algorithms could further increase the model efficiency.

Many studies are currently underway to provide individual assistance for health management through wearable devices and relevant mobile applications (Koushik et al., 2019; Liao et al., 2020). The current outcomes suggest that AI-based computational studies, combined with such innovative technologies, can not only evaluate sleep status in an individual manner but also promote early intervention by informing the risk of sleep disorders such as OSA. Those studies could ultimately contribute to personalized medicine.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Ethics statement

The studies involving human participants were reviewed and approved by Institutional Review Board of Ewha Womans University Mokdong Hospital (approval No. EUMC 2018–10-008). Owing to the retrospective and anonymized nature of this standard PSG database, the review board waived the need to obtain the patients' informed consent.

References

- Aboalayon, K. A. I., Faezipour, M., Almuhammadi, W. S., and Moslehpour, S. (2016). Sleep stage classification using EEG signal analysis: a comprehensive survey and new investigation. *Entropy* 18:272. doi: 10.3390/e18090272
- Acharya, U. R., Bhat, S., Faust, O., Adeli, H., Chua, E. C.-P., Lim, W. J. E., et al. (2015). Nonlinear dynamics measures for automated EEG-based sleep stage detection. *Eur. Neurol.* 74, 268–287. doi: 10.1159/000441975
- Altevogt, B. M., and Colten, H. R. (2006). *Sleep disorders and sleep deprivation: An unmet public health problem*. Washington, DC: National Academies Press, US.
- Berry, R., Quan, S., and Abreu, A. (2020). *The AASM manual for the scoring of sleep and associated events: Rules, terminology and technical specifications, version 2.6*. American Academy of Sleep Medicine, Darien.

Author contributions

CK, SA, and HL contributed to conception and design of the study. CK and SA contributed to construction of computational framework and interpretation of the data, and wrote the manuscript. HK, MD, HL, and SH contributed to data acquisition and data curation. HL contributed to data resources, interpretation of the data, funding acquisition, and substantively revised the article. All authors contributed to the manuscript and approved the submitted version.

Funding

This work was supported by grants from Institute of Information and communications Technology Planning and Evaluation (IITP) funded by the Korea government (MSIT) [No. RS-2022-00155966, Artificial Intelligence Convergence Innovation Human Resources Development (Ewha Womans University)], the Basic Science Research Program and Convergence Technology R&D Program for Human Augmentation (NRF-2019M3C1B8090803 and 2020R1A2C2013216), and by the BK21 FOUR (Fostering Outstanding Universities for Research) through the National Research Foundation of Korea (NRF) by the Korean government to HL. This work was also supported by the Basic Science Research Program through the NRF funded by the Ministry of Education (NRF-2020R1I1A1A01073605) to SA, and the Korea Health Technology R&D Project through the Korea Health Industry Development Institute funded by the Korean Ministry of Health & Welfare (HI19C1065) to HK.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fnins.2023.1059186/full#supplementary-material>

- Biswal, S., Sun, H., Goparaju, B., Westover, M. B., Sun, J., and Bianchi, M. T. (2018). Expert-level sleep scoring with deep neural networks. *J. Am. Med. Inform. Assoc.* 25, 1643–1650. doi: 10.1093/jamia/ocy131
- Bixler, E. O., Vgontzas, A. N., Lin, H. M., Ten Have, T., Rein, J., Vela-Bueno, A., et al. (2001). Prevalence of sleep-disordered breathing in women: effects of gender. *Am. J. Respir. Crit. Care Med.* 163, 608–613. doi: 10.1164/ajrccm.163.3.9911064
- Brunner, D., Vasko, R., Detka, C., Monahan, J., Reynolds, C. III, and Kupfer, D. (1996). Muscle artifacts in the sleep EEG: automated detection and effect on all-night EEG power spectra. *J. Sleep Res.* 5, 155–164. doi: 10.1046/j.1365-2869.1996.00009.x
- Campos-Beltrán, D., and Marshall, L. (2021). Changes in sleep EEG with aging in humans and rodents. *Pflügers Arch.* 473, 841–851. doi: 10.1007/s00424-021-02545-y
- Carrier, J., Land, S., Buysse, D. J., Kupfer, D. J., and Monk, T. H. (2001). The effects of age and gender on sleep EEG power spectral density in the middle years of life (ages 20–60 years old). *Psychophysiology* 38, 232–242. doi: 10.1111/1469-8986.3820232
- Chambon, S., Galtier, M. N., Arnal, P. J., Wainrib, G., and Gramfort, A. (2018). A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series. *IEEE Trans. Neural Syst. Rehabil. Eng.* 26, 758–769. doi: 10.1109/TNSRE.2018.2813138
- Choi, G. S., Yun, J. Y., Hwang, S., Kim, S. E., Kim, J.-Y., Im, C.-H., et al. (2021). Can Corticomuscular coherence differentiate between REM sleep behavior disorder with or without parkinsonism? *J. Clin. Med.* 10:5585. doi: 10.3390/jcm10235585
- Dong, H., Supratak, A., Pan, W., Wu, C., Matthews, P. M., and Guo, Y. (2018). Mixed neural network approach for temporal sleep stage classification. *IEEE Trans. Neural Syst. Rehabil. Eng.* 26, 324–333. doi: 10.1109/TNSRE.2017.2733220
- Enshaeifar, S., Kouchaki, S., Took, C. C., and Sanei, S. (2015). Quaternion singular spectrum analysis of electroencephalogram with application in sleep analysis. *IEEE Trans. Neural Syst. Rehabil. Eng.* 24, 57–67. doi: 10.1109/TNSRE.2015.2465177
- Fan, J., Sun, C., Long, M., Chen, C., and Chen, W. (2021). Eognet: a novel deep learning model for sleep stage classification based on single-channel eog signal. *Front. Neurosci.* 15:573194. doi: 10.3389/fnins.2021.573194
- Fogel, S. M., and Smith, C. T. (2011). The function of the sleep spindle: a physiological index of intelligence and a mechanism for sleep-dependent memory consolidation. *Neurosci. Biobehav. Rev.* 35, 1154–1165. doi: 10.1016/j.neubiorev.2010.12.003
- Ghasemzadeh, P., Kalbkhani, H., and Shayesteh, M. G. (2019). Sleep stages classification from EEG signal based on Stockwell transform. *IET Signal Process.* 13, 242–252. doi: 10.1049/iet-spr.2018.5032
- Gottlieb, D. J., and Punjabi, N. M. (2020). Diagnosis and management of obstructive sleep apnea. *JAMA* 323:1389. doi: 10.1001/jama.2020.3514
- Gul, A., Memtily, N., Aihemaiti, A., Abdulkadir, A., and Wushuer, P. (2018). Severity of obstructive sleep apnea (OSA) in 382 Han and Uyghur patients with sleep disorders. *Tradit. Med. Mod.* 1, 163–169. doi: 10.1142/S2575900018500106
- Hsu, Y.-L., Yang, Y.-T., Wang, J.-S., and Hsu, C.-Y. (2013). Automatic sleep stage recurrent neural classifier using energy features of EEG signals. *Neurocomputing* 104, 105–114. doi: 10.1016/j.neucom.2012.11.003
- Huang, C.-S., Lin, C.-L., Ko, L.-W., Liu, S.-Y., Su, T.-P., and Lin, C.-T. (2014). Knowledge-based identification of sleep stages based on two forehead electroencephalogram channels. *Front. Neurosci.* 8:263. doi: 10.3389/fnins.2014.00263
- Hussain, I., Hossain, M. A., Jany, R., Bari, M. A., Uddin, M., Kamal, A. R. M., et al. (2022). Quantitative evaluation of EEG-biomarkers for prediction of sleep stages. *Sensors* 22:3079. doi: 10.3390/s22083079
- Huttunen, R., Leppänen, T., Duce, B., Arnardottir, E. S., Nikkonen, S., Myllymaa, S., et al. (2022). A comparison of signal combinations for deep learning-based simultaneous sleep staging and respiratory event detection. *IEEE Trans. Biomed. Eng.* 70, 1704–1714. doi: 10.1109/TBME.2022.3225268
- Ioannides, A. A., Liu, L., and Kostopoulos, G. K. (2019). The emergence of spindles and K-complexes and the role of the dorsal caudal part of the anterior cingulate as the generator of K-complexes. *Front. Neurosci.* 13:814. doi: 10.3389/fnins.2019.00814
- Janiesch, C., Zschech, P., and Heinrich, K. (2021). Machine learning and deep learning. *Electron. Mark.* 31, 685–695. doi: 10.1007/s12525-021-00475-2
- Jarchi, D., Andreu-Perez, J., Kiani, M., Vysata, O., Kuchynka, J., Prochazka, A., et al. (2020). Recognition of patient groups with sleep related disorders using bio-signal processing and deep learning. *Sensors* 20:2594. doi: 10.3390/s20092594
- Kang, J. M., Cho, S.-E., Na, K.-S., and Kang, S.-G. (2021). Spectral power analysis of sleep electroencephalography in subjects with different severities of obstructive sleep apnea and healthy controls. *Nat. Sci. Sleep* 13, 477–486. doi: 10.2147/NSS.S295742
- Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization. *arXiv* 2014:6980. doi: 10.48550/arXiv.1412.6980
- Korkalainen, H., Aakko, J., Nikkonen, S., Kainulainen, S., Leino, A., Duce, B., et al. (2019). Accurate deep learning-based sleep staging in a clinical population with suspected obstructive sleep apnea. *IEEE J. Biomed. Health Inform.* 24, 2073–2081. doi: 10.1109/JBHI.2019.2951346
- Koushik, A., Amores, J., and Maes, P. (2019). Real-time smartphone-based sleep staging using 1-channel EEG. 2019 *IEEE 16th international conference on wearable and implantable body sensor networks*.
- Krauss, P., Schilling, A., Bauer, J., Tziridis, K., Metzner, C., Schulze, H., et al. (2018). Analysis of multichannel eeg patterns during human sleep: a novel approach. *Front. Hum. Neurosci.* 12:121. doi: 10.3389/fnhum.2018.00121
- Kryger, M. H., Roth, T., and Dement, W. C. (2017). *Principles and practice of sleep medicine*. 6th Edn. Philadelphia, PA: Elsevier.
- Kumari, U., Kora, P., Meenakshi, K., Swaraja, K., Padma, T., Panigrahy, A. K., et al. (2020). Feature extraction and detection of obstructive sleep apnea from raw eeg signal. *International Conference on Innovative Computing and Communications*.
- Kuo, C.-E., and Chen, G.-T. (2020). Automatic sleep staging based on a hybrid stacked LSTM neural network: verification using large-scale dataset. *IEEE Access* 8, 11837–11849. doi: 10.1109/ACCESS.2020.3002548
- Landolt, H.-P., Dijk, D.-J., Achermann, P., and Borbély, A. A. (1996). Effect of age on the sleep EEG: slow-wave activity and spindle frequency activity in young and middle-aged men. *Brain Res.* 738, 205–212. doi: 10.1016/S0006-8993(96)00770-6
- Li, Y., Wong, K. M., and de Bruin, H. (2012). Electroencephalogram signals classification for sleep-state decision—a Riemannian geometry approach. *IET Signal Process.* 6, 288–299. doi: 10.1049/iet-spr.2011.0234
- Liang, S.-F., Kuo, C.-E., Hu, Y.-H., Pan, Y.-H., and Wang, Y.-H. (2012). Automatic stage scoring of single-channel sleep EEG by using multiscale entropy and autoregressive models. *IEEE Trans. Instrum. Meas.* 61, 1649–1657. doi: 10.1109/TIM.2012.2187242
- Liao, Y., Zhang, M., Wang, Z., and Xie, X. (2020). *Tri-featurenet: an adversarial learning-based invariant feature extraction for sleep staging using single-channel EEG*. 2020 IEEE international symposium on circuits and systems (ISCAS).
- Loh, H. W., Ooi, C. P., Vicnesh, J., Oh, S. L., Faust, O., Gertych, A., et al. (2020). Automated detection of sleep stages using deep learning techniques: a systematic review of the last decade (2010–2020). *Appl. Sci.* 10:8963. doi: 10.3390/app10248963
- Long, X., Fonseca, P., Fossier, J., Haakma, R., and Aarts, R. M. (2013). Sleep and wake classification with actigraphy and respiratory effort using dynamic warping. *IEEE J. Biomed. Health Inform.* 18, 1272–1284. doi: 10.1109/JBHI.2013.2284610
- Lyons, M. M., Bhatt, N. Y., Pack, A. I., and Magalang, U. J. (2020). Global burden of sleep-disordered breathing and its implications. *Respirology* 25, 690–702. doi: 10.1111/resp.13838
- Malafeev, A., Laptev, D., Bauer, S., Omlin, S., Wierzbicka, A., Wichniak, A., et al. (2018). Automatic human sleep stage scoring using deep neural networks. *Front. Neurosci.* 12:781. doi: 10.3389/fnins.2018.00781
- Mander, B. A., Winer, J. R., and Walker, M. P. (2017). Sleep and human aging. *Neuron* 94, 19–36. doi: 10.1016/j.neuron.2017.02.004
- Marvaldi, M., Mallet, J., Dubertret, C., Moro, M. R., and Guessoum, S. B. (2021). Anxiety, depression, trauma-related, and sleep disorders among healthcare workers during the COVID-19 pandemic: a systematic review and meta-analysis. *Neurosci. Biobehav. Rev.* 126, 252–264. doi: 10.1016/j.neubiorev.2021.03.024
- Medley, M. L. (1980). Life satisfaction across four stages of adult life. *Int. J. Aging Hum. Dev.* 11, 193–209. doi: 10.2190/D4LG-ALJQ-8850-GYDV
- Mikkelsen, K., and De Vos, M. (2018). Personalizing deep learning models for automatic sleep staging. *arXiv* 2018:02645. doi: 10.48550/arXiv.1801.02645
- Mousavi, S., Afghah, F., and Acharya, U. R. (2019). SleepEEGNet: automated sleep stage scoring with sequence to sequence deep learning approach. *PLoS One* 14:e0216456. doi: 10.1371/journal.pone.0216456
- Panossian, L. A., and Avidan, A. Y. (2009). Review of sleep disorders. *Med. Clin. N. Am.* 93, 407–425. doi: 10.1016/j.mcna.2008.09.001
- Patil, S. P., Schneider, H., Schwartz, A. R., and Smith, P. L. (2007). Adult Obstructive Sleep Apnea. *Chest* 132, 325–337. doi: 10.1378/chest.07-0040
- Peters, K. R., Ray, L. B., Fogel, S., Smith, V., and Smith, C. T. (2014). Age differences in the variability and distribution of sleep spindle and rapid eye movement densities. *PLoS One* 9:e91047. doi: 10.1371/journal.pone.0091047
- Phan, H., Andreotti, F., Cooray, N., Chén, O. Y., and De Vos, M. (2018). *Automatic sleep stage classification using single-channel EEG: learning sequential features with attention-based recurrent neural networks*. In: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).
- Phan, H., Andreotti, F., Cooray, N., Chen, O. Y., and De Vos, M. (2019). SeqSleepNet: end-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging. *IEEE Trans. Neural Syst. Rehabil. Eng.* 27, 400–410. doi: 10.1109/tnsre.2019.2896659
- Pitkänen, H., Duce, B., Leppänen, T., Kainulainen, S., Kulkas, A., Myllymaa, S., et al. (2021). Gamma power of electroencephalogram arousal is modulated by respiratory event type and severity in obstructive sleep apnea. *IEEE Trans. Biomed. Eng.* 69, 1417–1423. doi: 10.1109/TBME.2021.3118229
- Sekkal, R. N., Bereksi-Reguig, F., Ruiz-Fernandez, D., Dib, N., and Sekkal, S. (2022). Automatic sleep stage classification: from classical machine learning methods to deep learning. *Biomed. Signal Process. Control* 77:103751. doi: 10.1016/j.bspc.2022.103751
- Sharma, M., Tiwari, J., and Acharya, U. R. (2021). Automatic sleep-stage scoring in healthy and sleep disorder patients using optimal wavelet filter bank technique with EEG signals. *Int. J. Environ. Res. Public Health* 18:3087. doi: 10.3390/ijerph18063087
- Simor, P., Horváth, K., Ujma, P. P., Gombos, F., and Bódizs, R. (2013). Fluctuations between sleep and wakefulness: wake-like features indicated by increased EEG alpha

- power during different sleep stages in nightmare disorder. *Biol. Psychol.* 94, 592–600. doi: 10.1016/j.biopsycho.2013.05.022
- Sors, A., Bonnet, S., Mirek, S., Vercueil, L., and Payen, J.-F. (2018). A convolutional neural network for sleep stage scoring from raw single-channel EEG. *Biomed. Signal Process. Control* 42, 107–114. doi: 10.1016/j.bspc.2017.12.001
- Stephansen, J. B., Olesen, A. N., Olsen, M., Ambati, A., Leary, E. B., Moore, H. E., et al. (2018). Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. *Nat. Commun.* 9:5229. doi: 10.1038/s41467-018-07229-3
- Supratak, A., Dong, H., Wu, C., and Guo, Y. (2017). DeepSleepNet: a model for automatic sleep stage scoring based on raw single-channel EEG. *IEEE Trans. Neural Syst. Rehabil. Eng.* 25, 1998–2008. doi: 10.1109/TNSRE.2017.2721116
- Supratak, A., and Guo, Y. (2020). *TinySleepNet: an efficient deep learning model for sleep stage scoring based on raw Single-Channel EEG*. In: 2020 42nd annual international conference of the IEEE engineering in medicine and Biology Society (EMBC).
- Tripathy, R. K., Ghosh, S. K., Gajbhiye, P., and Acharya, U. R. (2020). Development of automated sleep stage classification system using multivariate projection-based fixed boundary empirical wavelet transform and entropy features extracted from multichannel EEG signals. *Entropy* 22:1141. doi: 10.3390/e22101141
- Tzimourta, K. D., Tsilimbaris, A., Tzioukalia, K., Tzallas, A. T., Tsipouras, M. G., Astrakas, L. G., et al. (2018). EEG-based automatic sleep stage classification. *Biomed. J.* 7:1535. doi: 10.26717/BJSTR.2018.07.001535
- Verma, J. (2013). “One-way ANOVA: comparing means of more than two samples” in *Data analysis in management with SPSS software*. ed. J. P. Verma (Berlin: Springer), 221–254.
- Wei, R., Zhang, X., Wang, J., and Dang, X. (2018). The research of sleep staging based on single-lead electrocardiogram and deep neural network. *Biomed. Eng. Lett.* 8, 87–93. doi: 10.1007/s13534-017-0044-1
- Werth, E., Achermann, P., and Borbély, A. (1997). Fronto-occipital EEG power gradients in human sleep. *J. Sleep Res.* 6, 102–112. doi: 10.1046/j.1365-2869.1997.d01-36.x
- Willemsen, T., Van Deun, D., Verhaert, V., Vandekerckhove, M., Exadaktylos, V., Verbraecken, J., et al. (2013). An evaluation of cardiorespiratory and movement features with respect to sleep-stage classification. *IEEE J. Biomed. Health Inform.* 18, 661–669. doi: 10.1109/JBHI.2013.2276083
- Wu, H.-T., Talmon, R., and Lo, Y.-L. (2014). Assess sleep stage by modern signal processing techniques. *IEEE Trans. Biomed. Eng.* 62, 1159–1168. doi: 10.1109/TBME.2014.2375292
- Yuan, Y., Jia, K., Ma, F., Xun, G., Wang, Y., Su, L., et al. (2019). A hybrid self-attention deep learning framework for multivariate sleep stage classification. *BMC Bioinformatics* 20, 1–10. doi: 10.1186/s12859-019-3075-z
- Yue, H. J., Bardwell, W., Ancoli-Israel, S., Lored, J. S., and Dimsdale, J. E. (2009). Arousal frequency is associated with increased fatigue in obstructive sleep apnea. *Sleep Breath.* 13, 331–339. doi: 10.1007/s11325-009-0252-8
- Zhang, Y., Yang, Z., Lan, K., Liu, X., Zhang, Z., Li, P., et al. (2019). *Sleep stage classification using bidirectional lstm in wearable multi-sensor systems*. IEEE INFOCOM 2019-IEEE conference on computer communications workshops (INFOCOM WKSHPS).
- Zhao, D., Wang, Y., Wang, Q., and Wang, X. (2019). Comparative analysis of different characteristics of automatic sleep stages. *Comput. Methods Prog. Biomed.* 175, 53–72. doi: 10.1016/j.cmpb.2019.04.004
- Zhu, G., Li, Y., and Wen, P. (2014). Analysis and classification of sleep stages based on difference visibility graphs from a single-channel EEG signal. *IEEE J. Biomed. Health Inform.* 18, 1813–1821. doi: 10.1109/JBHI.2014.2303991