Standardized Image-Based Polysomnography Database and Deep Learning Algorithm for Sleep Stage Classification

Jaemin Jeong[1†], Wonhyuck Yoon[2†], Jeong-Gun Lee[1], Dongyoung Kim[1], Yunhee Woo[4], Dong-Kyu Kim[2,4,5*], and Hyun-Woo Shin[2,3,6,7,8,9*]

[1]School of Software, Hallym University, Chuncheon, Republic of Korea

[2]OUaR LaB, Inc, Seoul, Republic of Korea

[3]Obstructive Upper Airway Research (OUaR) Laboratory, Department of Pharmacology, Seoul National University College of Medicine, Seoul, Republic of Korea

[4]Institute of New Frontier Research, Division of Big Data and Artificial Intelligence, Chuncheon Sacred Heart Hospital, Hallym University College of Medicine, Chuncheon, Republic of Korea

[5]Department of Otorhinolaryngology-Head and Neck Surgery, Chuncheon Sacred Heart Hospital, Hallym University College of Medicine, Chuncheon, Republic of Korea

[6]Department of Biomedical Sciences, Seoul National University Graduate School, Seoul, Republic of Korea

[7]Cancer Research Institute, Seoul National University College of Medicine, Seoul, Republic of Korea

[8]Sensory Organ Research Institute, Seoul National University College of Medicine, Seoul, Republic of Korea

[9]Department of Otorhinolaryngology-Head and Neck Surgery, Seoul National University Hospital, Seoul, Republic of Korea.

[†]Jaemin Jeong and [†]Wonhyuck Yoon equally contributed to this work as the first author

***Co-Corresponding** Hyun-Woo Shin amd Dong-Kyu Kim

Hyun-Woo Shin, MD, PhD

Department of Pharmacology, Seoul National University College of Medicine and Department of Otorhinolaryngology-Head and Neck Surgery, Seoul National University Hospital, 103 Daehak-ro, Jongno-gu, Seoul (03080), Republic of Korea

Phone: 82-2-740-8285; Fax: 82-2-745-7996; E-mail: charlie@snu.ac.kr

Dong-Kyu Kim, MD, PhD

Division of Big Data and Artificial Intelligence, Department of Otorhinolaryngology-Head and Neck Surgery, Chuncheon Sacred Heart Hospital, Hallym University College of Medicine (24253), 77, Sakju-ro, Chuncheon-si, Gangwon-do, Republic of Korea

Phone: 82-33-240-5180; Fax: 82-33-241-2909; E-mail: doctordk@naver.com

Abstract

Study objectives: Polysomnography (PSG) scoring is labor intensive, subjective, and often ambiguous. Recently several deep learning (DL) models for automated sleep scoring have been developed, they are tied to a fixed amount of input channels and resolution. In this study, we constructed a standardized image-based PSG dataset in order to overcome the heterogeneity of raw signal data obtained from various PSG devices and various sleep laboratory environments.

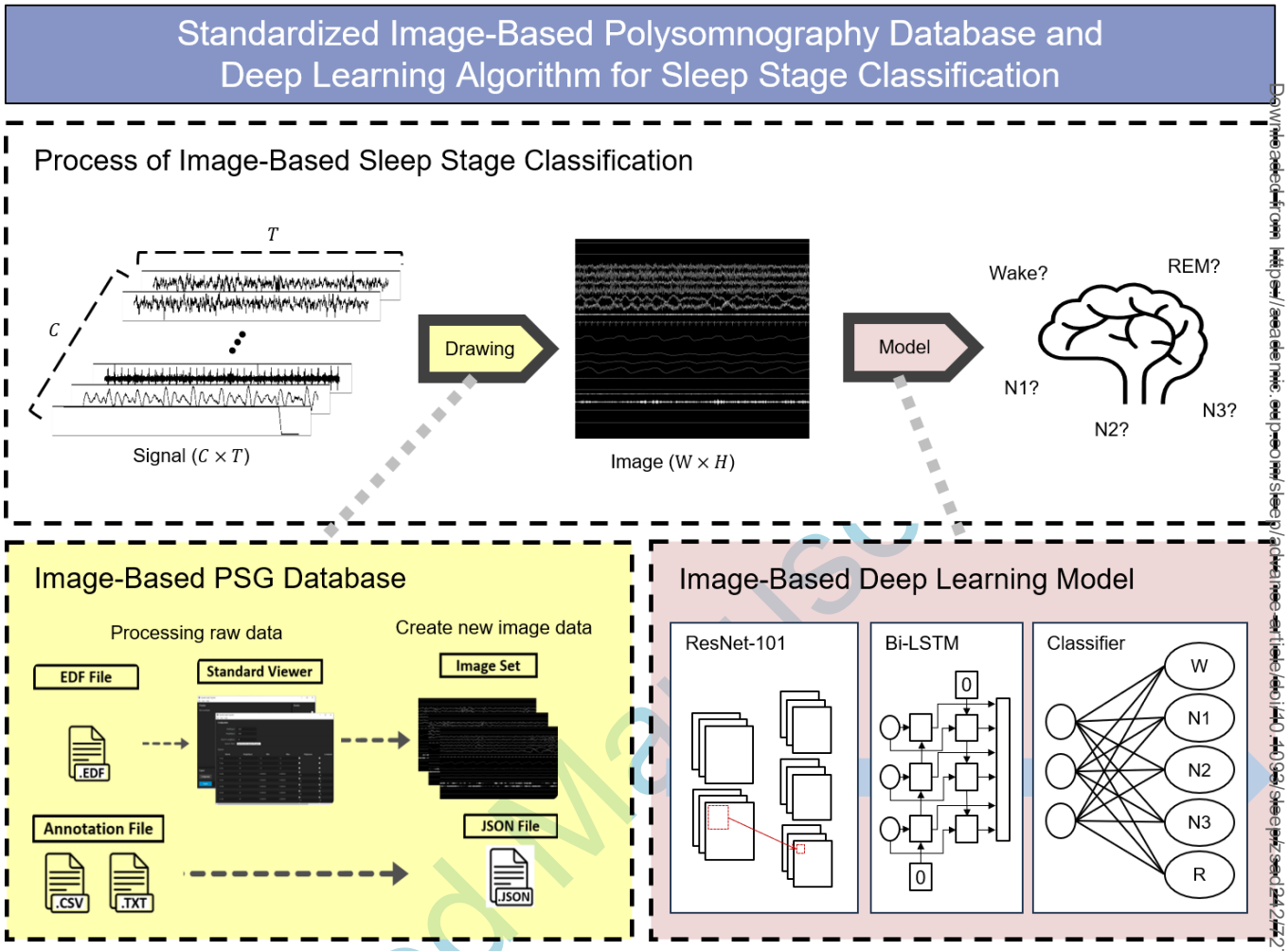
Methods: All individually exported European data format files containing raw signals were converted into images with an annotation file, which contained the demographics, diagnoses, and sleep statistics. An image-based DL model for automatic sleep staging was developed, compared with a signal-based model and validated in an external dataset

Results: We constructed 10,253 image-based PSG datasets using a standardized format. Among these, 7,745 diagnostic PSG data were used to develop our DL model. The DL model using the image dataset showed similar performance to the signal-based dataset for the same subject. The overall DL accuracy was greater than 80%, even with severe obstructive sleep apnea. Moreover, for the first time, we showed explainable DL in the field of sleep medicine as visualized key inference regions using Eigen-class activation maps. Furthermore, when a DL model for sleep scoring performs external validation, we achieved a relatively good performance.

Conclusion: Our main contribution demonstrates the availability of a standardized image-based dataset, and highlights that changing the data sampling rate or number of sensors may not require retraining, although performance decreases slightly as the number of sensors decreases.

Key Words— Sleep Stages, Polysomnography, Dataset, Deep Learning, Computer Neural Network

Graphical abstract



Statement of Significance

Polysomnography scoring is labor intensive and suffers from variability among scorers. Thus, various deep learning models have been developed to solve this problem; however, those based on signal data still have limitations for requiring to adjust when the sampling rate and amount of sensors change. Thus, although the performance of the DL model may be slightly decreased as the number of sensors decreased, we constructed a standardized image-based PSG dataset that does not require retraining regardless of the sampling rate and amount of sensor change.. It also confirmed the model generalization on the external validation. Therefore, this image-based PSG dataset can make the deep learning model more widely applicable compared to existing models that are tied to a fixed amount of input channels and resolution.

Introduction

Laboratory-based polysomnography (PSG), also known as the study of sleep, is currently the gold standard for the diagnosis of sleep-disordered breathing, such as obstructive sleep apnea (OSA). It uses a multiparametric measurement apparatus that records several physiological signals in parallel, including an electroencephalogram (EEG), electrocardiogram (ECG), electrooculogram (EOG), electromyogram (EMG), blood oxygenation, airflow, and respiratory effort. Thus, understanding the meaning of PSG data requires sleep scoring by a sleep technologist [1]. However, these expert-based sleep-stage systems have certain practical limitations. First, manual scoring methods are time-consuming and labor intensive, as sleep staging is still largely performed by clinicians in sleep clinics, even though the process is guided by well-established manuals. In addition, these time-consuming manual-scoring methods are unsuitable for processing large-scale data [2, 3]. Second, inter-scorer variability in sleep stage scoring is unavoidable [4-6]. A discrepancy in scoring is primarily detected in the combination of wake-N1, N1-N2, and N2-N3 between sleep centers [7]. Moreover, the inter-operator reliability of human expert scorers was found to be less than 0.8 as an inter-rater agreement, according to a recent meta-analysis study [8]. Third, there is a growing demand for self-monitoring of sleep status at home [9-11]. Therefore, automated sleep stage scoring using artificial intelligence has been thoroughly investigated in the field of sleep medicine. To date, several automatic sleep stage classification methods have been published. However, automatic sleep stage classification methods based on the signal dataset are currently not widely used because these are trained on a specific set of signal types and do not appropriately operate on datasets with different modalities. . Therefore, the sleep stage recognition technique requires manual examination of PSG by a sleep expert. To improve this, we propose a general deep learning (DL)-based sleep stage scoring system that uses an image-based PSG dataset.

Methods

This study was approved by the Institutional Review Board of Seoul National University Hospital (Seoul, Republic of Korea: No. C-2007-179-1143) and Chuncheon Sacred Heart Hospital, Hallym University College of Medicine (Chuncheon, Republic of Korea: No. 2021-03-005). A total of 10,253 PSG data points were used to build an image-based PSG dataset, which was constructed on the cloud server of the Korea National Information Society Agency (https://safezone.aihub.or.kr). Written informed consent was waived because the database was constructed in a deidentified manner.

Construction of an Image-Based Polysomnography Database

We collected PSG data from patients with sleep-disordered breathing from several sleep centers (Seoul National University Hospital, Hallym University Hospital, Seoul Sleep Center, and Lee & Hong Otorhinolaryngology Clinics), and constructed a new image-based PSG dataset. The CONSORT flow diagram for dataset construction and model development is given in Supplementary Figure 1. All numbers shown in the figure represent the number of PSG studies. Each sleep center recorded biosignals during sleep using computerized PSG devices such as Embla (Natus Medical, San Carlos, CA, USA) and Nox (Nox Medical, Reykjavik, Iceland).

All PSG data were scored by a sleep technologist, and then reviewed by another sleep technologist and a sleep specialist. The scoring process was conducted in accordance with the guidelines of the American Academy of Sleep Medicine (version 2.6). Especially, sleep stages were labeled in five classes: wake, non-rapid eye movement (non-REM) stage 1 (N1), non-REM stage 2 (N2), non-REM stage 3 (N3), and REM. After collecting the labeled PSG data, we extracted raw biosignals data into the European Data Format (EDF), and converted them into images using a developed program named "standard viewer". We also extracted all labels and statistical results (such as total sleep time, sleep latency, or apnea-hypopnea index (AHI)), and integrated them into JSON formatted files which were used as annotation files when DL models were trained (Figure 1A and Supplementary Figure 2). Each phase of the data construction procedure was conducted automatically by employing a robotic process automation tool.

During this procedure, data were de-identified by removing each patient's personal information such as name, patient number, and birth date, and the data of each patient were assigned a serial number. Figure 1B shows the format of the newly developed image-based PSG data. Each image had predefined areas for each channel, and the defined channel locations were the same for all images. These image-based data included the following 11 biosignals: EEG (C3–M2, C4–M1, O1–M2, O2–M1), EOG (E1–M2, E2–M1), chin EMG, ECG, airflow (oronasal thermistor), thoracic movement, abdominal movement, snoring sensor (audio volume), leg EMG (left and right leg), and oxygen saturation (85–100%, 40–100%). Thus, each image was 30 s long, with 11 channels of data corresponding to an epoch, and was displayed as a 1920 × 1080 high-resolution PNG file. The raw PSG data were also preprocessed with a 4th-order Butterworth low-pass filter, high-pass filter, and band-pass filter before being converted into an image. The cut-off frequencies recommended by the AASM scoring manual were used. Therefore, visual clues or features that can be found in biosignals, such as sleep spindles and K-complexes, are observed at a level similar to that used in actual clinical settings. Specifically, the change in the

physical value represented by a pixel varies depending on the signal type, as shown in Supplementary Table 1. Moreover, for cases where the measured data exceeded the set graph range, images were generated to allow graphs of the same signal to overlap. For example, the four graphs representing EEG signals can overlap with each other, while the two graphs representing EOG signals can also overlap with each other. Thus, during the image dataset construction, we repeatedly confirmed that all channels were appropriately clipped to the respective predefined areas while being converted into images. Additionally, it was checked if the examiner could identify overlapping biosignals by each one. Finally, quality validation of the constructed dataset was performed. Two sleep experts (one sleep technologist and one sleep specialist) from outside the medical center where the data were collected validated the quality of each set of converted images and an annotation file. They compared the signal graphs obtained by RemLogic and Noxturnal (PSG data viewers) with the images generated in this study (one epoch at a time). Both experts examined whether the images were properly converted and synchronized with the exported sleep-related events, as well as whether the exported events were correctly labeled (i.e., scored) based on the AASM scoring manual. During the examination, the data that were not synchronized or where the graph shapes did not match among themselves were either revised or excluded. Moreover, if there was even one epoch that did not match the original data, the whole PSG data, including that epoch, were excluded. Subsequently, the image dataset passed by both experts was used as the final dataset.

Neural Network Architecture for Image-Based Polysomnography Dataset

To date, convolutional neural network (CNN) with a bidirectional long short-term memory (Bi-LSTM) have been a common architecture for automated sleep-stage classification [12]. The learning process on DeepSleepNet was first completed in the CNN part without Bi-LSTM. Then, Bi-LSTM was added to the final feature map layer and trained to exploit sequential features between epochs [12, 13]. We initially utilized two different DL models: DeepSleepNet and ResNet101 combined with Bi-LSTM. After comparing the predictive performance between the two models, DeepSleepNet was selected as the primary DL model for the signal-based PSG dataset, given its proven efficacy for extracting features from raw signal data. In contrast, the ResNet101 2D convolutional layer was primarily designed for image-based data, which is a challenge to use on signal-based data. During the process, we attempted to replace all the 2D convolutional layers in ResNet101 with 1D convolutional layers. However, the desired results were not obtained as the training did not converge well; thus, it was concluded that image-based models and signal-based models require different approaches. Therefore, DeepSleepNet was finally chosen for the signal-based PSG dataset, which involved feature extraction through

large kernel sizes and strides for classification. Moreover, ResNet101 and Bi-LSTM was used for the image-based PSG dataset, given its suitability for extracting features from images through repeated iterations of small kernel sizes and strides. Figure 2 shows the overall flow of the DL algorithm using an image-based PSG dataset. First, we preprocessed the raw PSG signal data of size C × T (C: channel and T: time) according to the preprocessing procedure, such as sampling, filtering, and normalization. Subsequently, we drew the signal on the image (W: width and H: height), and the images were archived into a dataset. The newly developed image-based dataset was then used as an input to the DL model. ResNet101 were added to extract spatial information, and the Bi-LSTM was added to extract temporal information. Subsequently, we proceeded with training and testing processes. During the training process, we applied an image data augmentation scheme called LineOut and LineMix (Supplementary Figure 3, 4, and comments). Our code is publicly accessible at https://github.com/ai-for-sleep/sleep_stage_classification_for_image.

Evaluation and Comparison of the Deep Learning Model Performance According to the Dataset

We conducted a comparison between the two DL models with the same original PSG dataset to determine whether the performance of DL using an image-based dataset was better or comparable than that using a signal-based dataset. The overall flow for training and testing using the DL algorithm is shown (Supplementary Figure 5). First, we preprocessed raw PSG data by filtering and scaling. For filtering, we used high-path, low-path, and notch filters. To normalize the descriptor values, we added a MinMax scaler. Next, two different paths for generating signal- and image-based data were employed. PSG record filtering generates a dataset by excluding specific patients who do not have the required signals. The preprocessed PSG signal data were provided directly as inputs to the network per epoch unit. We employed DeepSleepNet as a model for automatic sleep-stage scoring using a signal-based PSG dataset. Specifically, to input an image-based PSG dataset, we first scaled the prepared image data to 224 × 224 and then created a test dataset by randomly selecting test data from the dataset (the random seed was fixed). At this time, the selected testing dataset was used for both signal- and image-based data training to consistently compare both methodologies. The dataset, excluding the test dataset, was then divided into five folds. Each fold was unique, and there was no common data among them. One-fold was designated as the validation dataset, whereas the remaining four folds were designated as the training dataset. The DL model validated every training epoch, and the training was stopped when there was no improvement in the accuracy performance for the last 10 training epochs. When all five models for the folds were completed, the probabilities of each model were extracted from the test dataset and averaged to obtain the final prediction

results. Finally, we assessed the model performance in terms of accuracy, micro-F1 score, and weighted-F1 score to determine the effect of sleep stage class imbalance in this dataset. The weighted accuracy was calculated as the average per-class stage accuracy.

Visualization of the Class Activation Map of the Images

We employed Eigen-class activation maps (CAM) to visualize a model decision made during inference processing [14-16]. Eigen-CAM is a visualization method that employs singular value decomposition. It computes and visualizes the principal components of the learned features/representations from convolutional layers. X_{in} and X_{out} represent the input and output feature maps, respectively. V_1 is the first eigenvector in the V matrix. $L_{Eigen-CAM}$ is visualized through the operation of the output feature map X_{out} and first eigenvector V_1 of the output feature map.

$$X_{out} = W \cdot X_{in} \quad (1)$$

$$X_{out} = U\Sigma V^T \quad (2)$$

$$L_{eigen-CAM} = X_{out} \cdot V_1 \quad (3)$$

Results

The composition of the dataset according to PSG type, sex, age, and PSG device is shown in Supplementary Table 2. This public dataset includes annotation files corresponding to each image file. In addition, in this dataset, every 10,253 PSG recordings contained each annotation file, including clinical medical information, demographic data, and sleep-related events. Specifically, sleep-related event labels include 1) five classes of sleep stages, 2) respiratory events (apnea, hypopnea, and desaturation), and 3) movement events (limb movement, periodic limb movement, and arousal events).

Performance of the Deep Learning Model Using the Image-Based Dataset

From the final image-based PSG dataset, we selected 7,745 patient data from diagnostic PSG (3,464 cases with the Nox A1 PSG system from Nox Medical, and 4,281 cases with the Embla N7000 series PSG system from Natus Medical) (Supplementary Table 3). To evaluate the performance of the DL algorithm, we split our data

into training, validation, and test datasets at a ratio of 80/10/10 by patients (Table 1). In our study, we utilized an ensemble five-fold model for calculating the F1-score and generating the Confusion matrix. Furthermore, we employed a trained model using datasets divided at a ratio of 80/10/10 by patients for the remaining experiments. It is important to note that the same test dataset was used in all experiments. The diagnostic PSG cases and overall epoch volume according to the device or the institute are presented in Supplementary Table 4. For the five-stage classification of sleep stages (wake/N1/N2/N3/REM), we utilized ResNet101 and LSTM to extract the local spatiotemporal characteristics of 30-second PSGs. Our DL model trained with an image-based PSG dataset achieved an epoch-by-epoch accuracy of 82.91%, micro-F1 score of 82.90%, and weighted-F1 score of 82.76%. Additionally, when AHI values were converted into standard clinical categories of normal, mild, moderate, and severe diseases, our DL model trained with the image-based PSG dataset showed an overall accuracy of 86.84% for normal, 86.62% for mild, 84.44% for moderate, and 80.74% for severe diseases (Table 2 and Supplementary Figure 6). The sleep metrics trends exhibit a declining pattern in accuracy as the severity level increases [39, 40].

Comparison of Deep Learning Performance between Image-Based and Signal-Based Datasets

To identify whether the image-based PSG dataset could be useful for improving the performance of the DL algorithm, we compared the model performance of the image-based and signal-based datasets. The model performance obtained using the image-based PSG dataset for the five sleep classes was 82.91% overall accuracy, 82.90% macro F1-score, and 82.76% weighted F1-score, whereas that obtained using the signal-based PSG dataset was 81.88% for overall accuracy, 80.89% for macro F1-score, and 81.62% for weighted F1-score. The performance of the DL model obtained using the image-based PSG dataset was similar to the performance of the DL model obtained using the signal-based PSG dataset (Table 3). For the model performance of all the tested epochs, a confusion matrix was generated (Figure 3). When considering all epochs, the model using the image-based dataset scored the Wake, N1, N2, N3, and REM stages correctly 90%, 62%, 84%, 83%, and 89% of the time, respectively. Meanwhile, the model using the signal-based dataset correctly scored Wake, N1, N2, N3, and REM stages 86%, 68%, 80%, 83%, and 91% of the time, respectively. To compare the prediction performance between the two datasets, we used the area under the precision-recall curve (AUPRC) and the area under the receiver operating characteristic curve (AUROC) because the statistical comparison of overall accuracy scores was difficult between the two datasets (Figure 4). Although we could not access statistical differences, our findings showed that both AUPRC and AUROC were visually similar in each sleep stage

between the signal- and image-based datasets. Next, we evaluated the difference in DL model performance based on the PSG device (Table 3). Various studies have previously reported that when a DL model for sleep scoring performs external validation, performance decreases significantly. Similarly, in this study, when training and testing were performed using the dataset obtained from different PSG devices, the DL model using the signal-based dataset showed a significant decrease (4–8%), whereas the DL model using the image-based dataset revealed a relatively smaller decrease in its performance (approximately 4%).

Visualization of Major Determining Area to Score Sleep Stage in the Deep Learning Model

To determine which area was the deciding component for classification in the DL model using the image-based PSG dataset, we used Eigen-CAM. The final convolution layer contains spatial information indicating discriminative regions to make classifications and generates a spatial heatmap from the activations of the previous convolutional layer. Figure 5A shows these discriminative parts representing the image data of each sleep stage. In addition, we visualized the discriminative region for sleep staging by averaging each Eigen-CAM, which consisted of 10,000 images, to demonstrate where the model usually focused on each class (Figure 5B). Additionally, the discriminative regions for sleep staging obtained by averaging each Eigen-CAM are very similar in each sleep stage (Supplementary Figure 7). Because the image formats, such as the height, width, or position of the channels, are always the same, the CAM can visualize which parts of the images contribute to the model's decisions while maintaining consistency. Thus, discriminative information could provide the sleep clinician with an additional message, such as the focusing area for manual sleep staging. Interestingly, we investigated how each channel affects DL model performance during sleep stage classification. When only one channel in each image-based dataset was cleared, we found that EEG, followed by EOG, was the most important predictor of the DL model for sleep staging (Figure 6A). Meanwhile, when only one channel is left, we discovered that EEG and EOG values are crucial in reducing error and improving prediction accuracy when included in the DL model using an image-based dataset (Figure 6B). Similar findings were also obtained when we investigated the channel effects on the DL model performance according to sleep stages.

External Validation of the Image-Based Deep Learning Model with SHHS dataset

To evaluate the generalization ability of our image data format, we conducted an external validation with an open-access dataset. We utilized 2,652 PSG data from Sleep Heart Health Study (SHHS) Visit 2 dataset as the external data and converted them to our image data format. Since each PSG data from SHHS Visit 2 dataset has fewer channels compared to our dataset (e.g., it has only 2 EEG channels, whereas our data has 4 EEG

channels), we considered two approaches to address this issue. The first approach is called the "blanked" method, which involves treating the areas corresponding to non-existent channels (such as acceleration, O1-M2 and O2-M1 from EEG, and snoring signals etc.) as blanks (Figure 7A). The second approach referred to as the "duplicated" method. If a signal does not exist, the image is drawn using an available signal of the same type as the non-existent one (Figure 7B). We replaced the missing 2 EEG channels O1-M2 and O2-M1 with C3-M2 and C4-M1, and for the breathing-related signals, we duplicated Airflow (thermistor, refer to SHHS dataset website: https://sleepdata.org/datasets/shhs/pages/11-montage-and-sampling-rate-information-shhs2.md) signal to draw non-existent nasal pressure signal. Although both methods yielded similar performance, the "duplicated" method demonstrated slightly better results. With this external dataset, which was not used for training our image-based model, we achieved relatively good performance for Wake, N2, N3, and REM stages (Figure 7).

Discussion

To date, several researchers have attempted to develop an automatic classification of sleep stages using DL models, but only a few studies have demonstrated its effectiveness [17-22]. Therefore, we attempted to construct a standardized image-based database that is more effective for artificial intelligence learning, comprising a complete dataset that consists of the relevant polysomnographic and demographic data for all patients covered by the database. This database also includes polysomnographic data obtained from multiple sleep centers using two different types of PSG devices. To test our proposed PSG database, we developed a DL algorithm for automated sleep stage classification using the image-based PSG dataset and compared its performance with that of the DL algorithm using the raw signal-based PSG dataset. Although the processing of raw signals into images may impair the learning process because most of the image area is expressed as a black background, our results demonstrate that the automated sleep stage classification trained on this standardized public dataset could achieve a similar performance compared to those based on the raw signal. Additionally, our standardized image-based database showed similar performance in the external validation test.

Sleep-EDF, Sleep-EDF [Expanded], the Montreal Archive of sleep studies, the Sleep Heart Health Study collection, and the Massachusetts General Hospital (MGH) sleep laboratory database are well-known PSG datasets that exist, regardless of their public availability [24-27]. Of these datasets, the MGH database was the largest, with PSG recordings from 10,000 subjects. The main goal of constructing our PSG database was to create an optimized dataset for DL. Recent studies have shown that deep neural networks that use spectrogram representations of EEG segments outperform those that use raw EEG segments in terms of accuracy [19, 20,

28]. Similar to these reports, sleep technologists still scored the PSG data for visual pattern recognition. However, the spectrogram data-based methods need to extract a 2D amplitude signal, having time and frequency as its dimensions for each channel. Moreover, our image-based method contains all channels in one gray image. It means that our image-based data is much more efficient and even when using explainable artificial intelligence methods such as CAM, our image-based data are more intuitive to interpret than spectrogram data. Moreover, our image-based PSG dataset, which contained more than 10,000 subjects, is one of the largest PSG datasets linked to relevant clinical data. Similar to the MGH dataset, our image-based PSG database consisted of a mixture of diagnostic and titration protocols. Specifically, our database integrates 11 biosignals from PSG and annotated files for patient demographics and sleep statistics. Furthermore, when we considered clinical PSG scoring, we realized that high-quality PSG scoring of healthy people was insufficient. In this regard, our image-based database not only has the largest size but also has a greater proportion of unhealthy subjects. Interestingly, unlikely the signal-based PSG dataset, our image-based PSG dataset could add a unique direction to real-world applications. As a standardized form of the PSG dataset expressed as an image file is proposed in this study, it is possible to overcome the heterogeneity of PSG recordings from different sleep centers. Specifically, when training and test processes were performed using data obtained from different PSG devices, it was revealed that the DL model with the image-based PSG dataset outperformed the one with the signal-based dataset.

Generally, DL is known as a black-box model, which makes it difficult to provide a logical basis for output results [29, 30]. DL has recently been studied in the field of explainable artificial intelligence [31-33]. However, this is still a difficult problem. Thus, as a black box model with a multilayer nonlinear structure, deep neural networks are often criticized for being non-transparent, with their predictions being untraceable by humans [34]. In automatic sleep-staging systems, black-box skepticism remains one of the main questions regarding their clinical value and adoption because sleep stages are often ambiguous, and different human experts tend to disagree to some extent [35, 36]. However, when developing a model using image-based learning, several methods have been introduced to determine which part influences the output decision. Among these methods, CAM technology provides visualization of the final decision [15, 37, 38]. Therefore, it can learn based on images and obtain visualization information regarding the output of the model during inference, allowing the physician to effectively utilize the evaluation results. For these reasons, our approach has the advantage of being able to use existing SOTA image recognition models regardless of the input sampling rate. Importantly, when compared to research that focuses on frequency domain data, our approach allows medical staffs to confirm the reasoning behind the classification performed by deep learning models through an Eigen CAM on the time axis

of the PSG data they are viewing. These findings revealed that the DL model evaluated wake status mainly by respiration or chest/abdominal movements rather than EEG, whereas the inference of REM sleep depended on EOG and chin EMG rather than EEG (Supplementary Figure 7). DL methods using signal-based data have an insufficient effect on the extraction of EEG frequency information, resulting in a poor classification performance, particularly for N1 and REM. Thus, our image-based dataset provides an opportunity to improve the performance of N1 and REM classification. However, when the relevance of the entire biosignal channel was evaluated, omitting the EEG channel, followed by the EOG channel, significantly decreased the accuracy of our DL model (Figure 6). The remaining biosignal channels had a minimal effect on the accuracy of the DL model.

In addition, our standardized image-based database offers several advantages. Because PSG data are multichannel, it is essential to build a multichannel model when using the original raw signal data. Therefore, it may be difficult to learn multichannel features effectively because of the increase in complexity; however, with image-based data, multichannel signal data can be captured as a single-channel image for learning, lowering the complexity of DL models. Moreover, although higher frequencies could not detect in the image-based PSG dataset, it does not require retraining when the sampling rate and the number of sensors change. This makes the proposed dataset more widely applicable compared to existing ones, which are tied to fixed amounts of input channels at specific resolutions. Finally, we tried to test for external validation regarding image-based datasets using the raw signal of SHHS. On the external validation, the DL algorithm using the image-based dataset showed relatively good performance for each sleep stage; thus, we could confirm the model generalization based on the image dataset.

In conclusion, we constructed an image-based PSG database, one of the largest PSG databases linked to relevant clinical data. Thus, this database consists of all 11 biosignal waveform images, except the numeric parameters, as one image file, and each dataset also included patient demographics and sleep statistics. Additionally, the DL algorithm using the image-based PSG dataset achieved a similar performance, compared to those using this signal-based PSG dataset. When we performed the external validation, we confirmed the reliable results. Therefore, these findings indicate that our standardized image-based dataset may be effective for the development of DL-based automatic classification of sleep stages.

Data Availability

Although the full dataset cannot be made publicly available because of legal restrictions imposed by the Korean government in relation to the Personal Information Protection Act, if some investigators wish to use this image-based PSG data, they could access it after obtaining the relevant permit from the Korean National Information Society Agency (https://eng.nia.or.kr/site/nia_eng/main.do) has been granted.

Funding

This research was supported by the SNUH Research Fund (grant no 0320202090), AI-Bio Research Grant through Seoul National University and was a part of the "AI Dataset Project" (aihub.or.kr), funded by the Ministry of Science & ICT and National Information Society Agency, Republic of Korea.

Financial disclosure

H.-W. S. is an inventor on patent applications submitted by Seoul National University related to an image-based polysomnography dataset and its application. H.-W. S. is a founder of OUaR LaB, Inc., serves on the Board of Directors and as a chief executive officer for OUaR LaB, Inc., and owns OUaR LaB Stock, which are subject to certain restrictions under university policy. D.-K. K. also serves on the Board of Directors for OUaR LaB, Inc. and owns OUaR LaB Stock, which are subject to certain restrictions under university policy. Within the last one year, D.-K. K. and H.-W. S. have been an advisor to LG electronics. These arrangements have been reviewed and approved by Seoul National University or Hallym university in accordance with its conflict of interest policies. W.-H. Y. owns OUaR LaB Stock.

Non-financial disclosure

Authors have no conflict of non-financial interest to declare

References

[1] Berry RB, et al. Rules for scoring respiratory events in sleep: update of the 2007 AASM Manual for the Scoring of Sleep and Associated Events. Deliberations of the sleep apnea definitions task force of the American Academy of Sleep Medicine. J. Clin. Sleep Med. 2012; 8 (5): 597-619. doi: 10.5664/jcsm.2172.

[2] Chattu VK et al. The global problem of insufficient sleep and its serious public health implications. Healthcare (Basel). 2018; 7 (1): 1. doi: 10.3390/healthcare7010001.

[3] Phan H and Mikkelsen K, Automatic sleep staging of EEG signals: recent development, challenges, and future directions. Physiol. Meas. 2022; 43 (4). doi: 10.1088/1361-6579/ac6049.

[4] Collop NA. Scoring variability between polysomnography technologists in different sleep laboratories. Sleep Med. 2002; 3 (1): 43-47. doi: 10.1016/s1389-9457(01)00115-0.

[5] Loredo JS et al. Night-to-night arousal variability and interscorer reliability of arousal measurements. Sleep. 1999; 22 (7): 916-920. doi: 10.1093/sleep/22.7.916.

[6] Norman, RG et al. Interobserver agreement among sleep scorers from different centers in a large dataset. Sleep. 2000; 23 (7): 901-908. doi: 10.1093/sleep/23.7.1e.

[7] Deng S et al. Interrater agreement between American and Chinese sleep centers according to the 2014 AASM standard. Sleep Breath. 2019; 23 (2): 719-728. doi: 10.1007/s11325-019-01801-x.

[8] Lee YJ et al. Interrater reliability of sleep stage scoring: a meta-analysis,, J. Clin. Sleep Med. 2022; 18 (1): 193-202. doi: 10.5664/jcsm.9538.

[9] Yoon DW and Shin HW. Sleep tests in the non-contact era of the COVID-19 pandemic: home sleep tests versus in-laboratory polysomnography, Clin. Exp. Otorhinolaryngol. 2020; 13 (4): 318-319. doi: 10.21053/ceo.2020.01599.

[10] Di Pumpo M, et al. Multiple-access versus telemedicine home-based sleep apnea testing for obstructive sleep apnea (OSA) diagnosis: a cost-minimization study. Sleep Breath. 2022; 26 (4): 1641-1647. doi: 10.1007/s11325-021-02527-5.

[11] Kim, RD, et al. An economic evaluation of home versus laboratory-based diagnosis of obstructive sleep apnea. Sleep. 2015; 38 (7): 1027-1037. doi: 10.5665/sleep.4804.

[12] Faust O, et al. A review of automated sleep stage scoring based on physiological signals for the new millennia. Comput. Methods Programs Biomed. 2019; 176: 81-91. doi: 10.1016/j.cmpb.2019.04.032.

[13] Supratak A, et al. DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG. IEEE Trans. Neural Syst. Rehabil. Eng. 2017; 25 (11): 1998-2008. doi: 10.1109/TNSRE.2017.2721116.

[14] M. B. Muhammad and M. Yeasin. Eigen-CAM: Class Activation Map using Principal Components. International Joint Conference on Neural Networks (IJCNN). 2020: 1-7. doi: 10.1109/IJCNN48605.2020.9206626. [15] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva and A. Torralba. Learning Deep Features for Discriminative Localization. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016: 2921-2929. doi: 10.1109/CVPR.2016.319.[16] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. IEEE International Conference on Computer Vision (ICCV). 2017: 618-626. doi: 10.1109/ICCV.2017.74..

[17] Stephansen JB, et al. Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. Nat. Commun. 2018; 9 (1): 5229. doi: 10.1038/s41467-018-07229-3.

[18] Patanaik A, et al. An end-to-end framework for real-time automatic sleep stage classification. 2018; Sleep. 41 (5). doi: 10.1093/sleep/zsy041.

[19] Biswal S, et al. Expert-level sleep scoring with deep neural networks. J. Am. Med. Inform. Assoc. 2018; 25 (12): 1643-1650. doi: 10.1093/jamia/ocy131.

[20] Zhang L, et al. Automated sleep stage scoring of the Sleep Heart Health Study using deep neural networks. Sleep. 2019; 42 (11). doi: 10.1093/sleep/zsz159.

[21] Abou Jaoude M, et al. Expert-level automated sleep staging of long-term scalp electroencephalography recordings using deep learning. Sleep. 2020; 43 (11). doi: 10.1093/sleep/zsaa112.

[22] Xu Z, et al. Sleep stage classification based on multi-centers: comparison between different ages, mental health conditions and acquisition devices. Nat. Sci. Sleep. 2022; 14: 995-1007. doi: 10.2147/NSS.S355702.

[23] Lee AY, et al. Multicenter, head-to-head, real-world validation study of seven automated artificial intelligence diabetic retinopathy screening systems. Diabetes Care. 2021; 44 (5): 1168-1175. doi: 10.2337/dc20-1877.

[24] Goldberger AL, et al. PhysioBank, PhysioToolkit, and Physionet: components of a new research resource for complex physiologic signals. Circulation. 2000; 101 (23): E215-E220. doi: 10.1161/01.cir.101.23.e215.

[25] O'Reilly C, et al. Montreal Archive of Sleep Studies: an open-access resource for instrument benchmarking and exploratory research. J. Sleep Res. 2014; 23 (6): 628-635. doi: 10.1111/jsr.12169.

[26] Quan SF, et al. The Sleep Heart Health Study: design, rationale, and methods. Sleep. 1997; 20 (12): 1077-1085.

[27] Kemp B, et al. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG. IEEE Trans. Bio Med. Eng. 2000; 47 (9): 1185-1194. doi: 10.1109/10.867928.

[28] Li C, et al. A deep learning method approach for sleep stage classification with EEG spectrogram, Int. J. Environ. Res. Public Health. 2022; 19 (10): 6322. doi: 10.3390/ijerph19106322.

[29] Sarker IH. Deep learning: A comprehensive overview on techniques, taxonomy, applications and research directions. SN Comput. Sci. 2021; 2 (6): 420. doi: 10.1007/s42979-021-00815-1.

[30] Alzubaidi L, et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J. Big Data. 2021; 8(1): 53. doi: 10.1186/s40537-021-00444-8.

[31] Payrovnaziri SN, et al. Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. J. Am. Med. Inform. Assoc. 2020; 27(7): 1173-1185. doi: 10.1093/jamia/ocaa053.

[32] Giuste F, et al. Explainable artificial intelligence methods in combating pandemics: A systematic review. 2022; IEEE Rev. Biomed. Eng. PP. doi: 10.1109/RBME.2022.3185953.

[33] Lauritsen SM, et al. Explainable artificial intelligence model to predict acute critical illness from electronic health records. Nat. Commun. 2020; 11(1): 3852. doi: 10.1038/s41467-020-17431-x.

[34] Rajkomar A, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit. Med. 2018; 1(1): 18. doi: 10.1038/s41746-018-0029-1.

[35] Guillot A and Thorey V. RobustSleepNet: transfer learning for automated sleep staging at scale. IEEE Trans. Neural Syst. Rehabil. Eng. 2021; 29: 1441-1451. doi: 10.1109/TNSRE.2021.3098968.

[36] Danker-Hopfe H, et al. Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard. J. Sleep Res. 2009; 18(1): 74-84. doi: 10.1111/j.1365-2869.2008.00700.x.

[37] Ko YC, et al. Deep learning assisted detection of glaucomatous optic neuropathy and potential designs for a generalizable model. PLOS ONE. 2020; 15(5): e0233079. doi: 10.1371/journal.pone.0233079.

[38] Kim HG, et al. Improvement diagnostic accuracy of sinusitis recognition in paranasal sinus X-ray using multiple deep learning models. Quant. Imaging Med. Surg. 2019; 9(6): 942-951. doi: 10.21037/qims.2019.05.15.

[39] Sridhar, N., Shoeb, A., Stephens, P., Kharbouch, A., Shimol, D.B., Burkart, J., Ghoreyshi, A. and Myers, L., 2020. Deep learning for automated sleep staging using instantaneous heart rate. NPJ digital medicine, 3(1), p.106. doi: 10.1038/s41746-020-00337-9.

[40] Younes, M., Raneri, J. and Hanly, P., 2016. Staging sleep in polysomnograms: analysis of inter-scorer variability. Journal of Clinical Sleep Medicine, 12(6), pp.885-894. doi: 10.5664/jcsm.5894.

Figure legends

Figure 1. (A) Example of an image-based polysomnography data file (B) Schematic flow of construction of polysomnography database

Figure 2. Framework for automatic sleep-stage classification: after performing preprocessing such as sampling, filtering, and normalization on the input raw signal dataset, a standardized image dataset was created. Next, to test the five-class sleep-stage scoring, the standardized image dataset was added into the deep learning model, which combines the bidirectional long short-term memory network (Bi-LSTM) and fully connected layer with the convolutional neural network.

Figure 3. Confusion matrix of ensemble five-fold models for automatic sleep-stage classification (A) Signal-based PSG database, (B) Image-based PSG database.

Figure 4. Comparison of the predictive performance between signal- and image-based datasets. The performance was evaluated using both the area under the precision recall curve (AUPRC) and are under the receiver operating characteristic curve (AUROC). In this experiment, we use the test dataset split by patients.

Figure 5. (A) Class activation maps (CAM) for each sleep stage: a map is generated for each class of the network by obtaining the weighted sum of the last convolutional features using the fully connected layer weights. (B) Average of CAM for each sleep stage: it is obtained from about 10,000 images to see where the model usually focuses on each class. In this experiment, we use the test dataset split by patients.

Figure 6. Evaluation of the weighted-F1 score to understand the effect of each biosignal on the performance of the DL model: (A) when only one specific channel is erased, (B) when only one specific channel is left, (C) performance when one specific channel is erased in each sleep stage, and (D) performance when one specific channel is left in each sleep stage. In this experiment, we use the test dataset split by patients.

Figure 7. Confusion matrix of ensemble five-fold models for external validation of SHHS dataset. (A) Confusion matrix evaluation to convert non-existent channels to blank images, (B) Confusion matrix evaluation to convert non-existent channels to duplicate images, (C) Table representing the F1-score for A and B.

List of Tables

Table 1. Profile of datasets according to the severity of obstructive sleep apnea. Diagnostic PSG data were divided into training, validation, and test data in ratio of 80:10:10. The upper value in each cell represents the number of epochs, and the value in parentheses represents the proportion of each sleep stage according to the severity of sleep apnea.

Dataset	Severity	Wake	N1	N2	N3	REM	Total
Training	Normal	115,523 (19%)	51,967 (8%)	222,849 (37%)	116,409 (19%)	103,701 (17%)	610,449 (100%)
	Mild	134,191 (20%)	65,712 (9%)	249,704 (37%)	112,961 (17%)	113,279 (17%)	675,847 (100%)
	Moderate	190,007 (21%)	104,538 (12%)	326,076 (36%)	135,978 (15%)	144,842 (16%)	901,441 (100%)
	Severe	621,320 (26%)	460,710 (19%)	795,038 (32%)	244,743 (10%)	313,701 (13%)	2,435,512 (100%)
	Total	1,061,041 (23%)	682,927 (15%)	1,593,667 (35%)	610,091 (13%)	675,523 (14%)	**4,623,249** (100%)
Validation	Normal	16,511 (21%)	7,512 (9%)	29,095 (36%)	14,191 (18%)	12,773 (16%)	80,082 (100%)
	Mild	13,596 (18%)	7,271 (10%)	26,332 (36%)	13,998 (19%)	12,101 (17%)	73,298 (100%)
	Moderate	22,302 (20%)	13,034 (11%)	41,229 (37%)	16,990 (15%)	18,823 (17%)	112,378 (100%)
	Severe	79,029 (26%)	61,499 (20%)	97,725 (32%)	28,374 (9%)	37,746 (13%)	304,373 (100%)
	Total	131,438	89,316	194,381	73,553	81,443	**570,131**

		(23%)	(15%)	(35%)	(13%)	(14%)	(100%)
Test	Normal	10,791	4,901	19,994	9,866	9,526	55,078
		(20%)	(9%)	(36%)	(18%)	(17%)	(100%)
	Mild	19,434	8,471	34,142	14,373	14,256	90,676
		(21%)	(9%)	(38%)	(16%)	(16%)	(100%)
	Moderate	24,057	12,949	41,308	16,253	18,986	113,553
		(21%)	(11%)	(36%)	(14%)	(17%)	(100%)
	Severe	80,373	57,537	104,671	33,208	41,475	317,264
		(26%)	(18%)	(33%)	(10%)	(13%)	(100%)
	Total	134,655	83,858	200,115	73,700	84,243	**576,571**
		(23%)	(15%)	(35%)	(13%)	(14%)	(100%)

Table 2. Performance of deep learning model using the image-based dataset according to the severity of obstructive sleep apnea.

Severity	Accuracy (%)	Macro F1-score (%)	Weighted F1-score (%)
Normal	86.84	83.60	86.54
Mild	86.62	83.37	86.30
Moderate	84.44	81.98	84.11
Severe	80.74	80.68	80.66

The accuracy was calculated as the ratio of the number of correctly classified instances to the total number of instances in the dataset. The macro F1-score, which is the unweighted average of the F1-score for each class, was used as an evaluation metric.

The weighted F1-score accounts for the class imbalance in the dataset; it calculates the F1-score of each class weighted by its frequency and was also used as an evaluation metric.

Table 3. Comparison of deep learning model performance according to the origin of dataset and polysomnography device.

Training data	Test data	Signal-based dataset			Image-based dataset		
		Accuracy (%)	Macro F1-score (%)	Weighted F1-score (%)	Accuracy (%)	Macro F1-score (%)	Weighted F1-score (%)
All	All	81.48	80.89	81.62	82.91	82.90	82.76
Embla	Nox	74.13	72.75	74.19	78.63	77.53	78.68
Embla	Embla	82.83	82.06	82.88	83.35	82.43	83.24
Nox	Embla	76.40	74.79	76.29	77.11	74.77	76.16
Nox	Nox	80.55	79.98	80.69	81.94	80.79	81.71

The accuracy was calculated as the ratio of the number of correctly classified instances to the total number of instances in the dataset. The macro F1-score, which is the unweighted average of the F1-score for each class, was used as an evaluation metric.

The weighted F1-score accounts for the class imbalance in the dataset; it calculates the F1-score of each class weighted by its frequency and was also used as an evaluation metric.

Figure 1

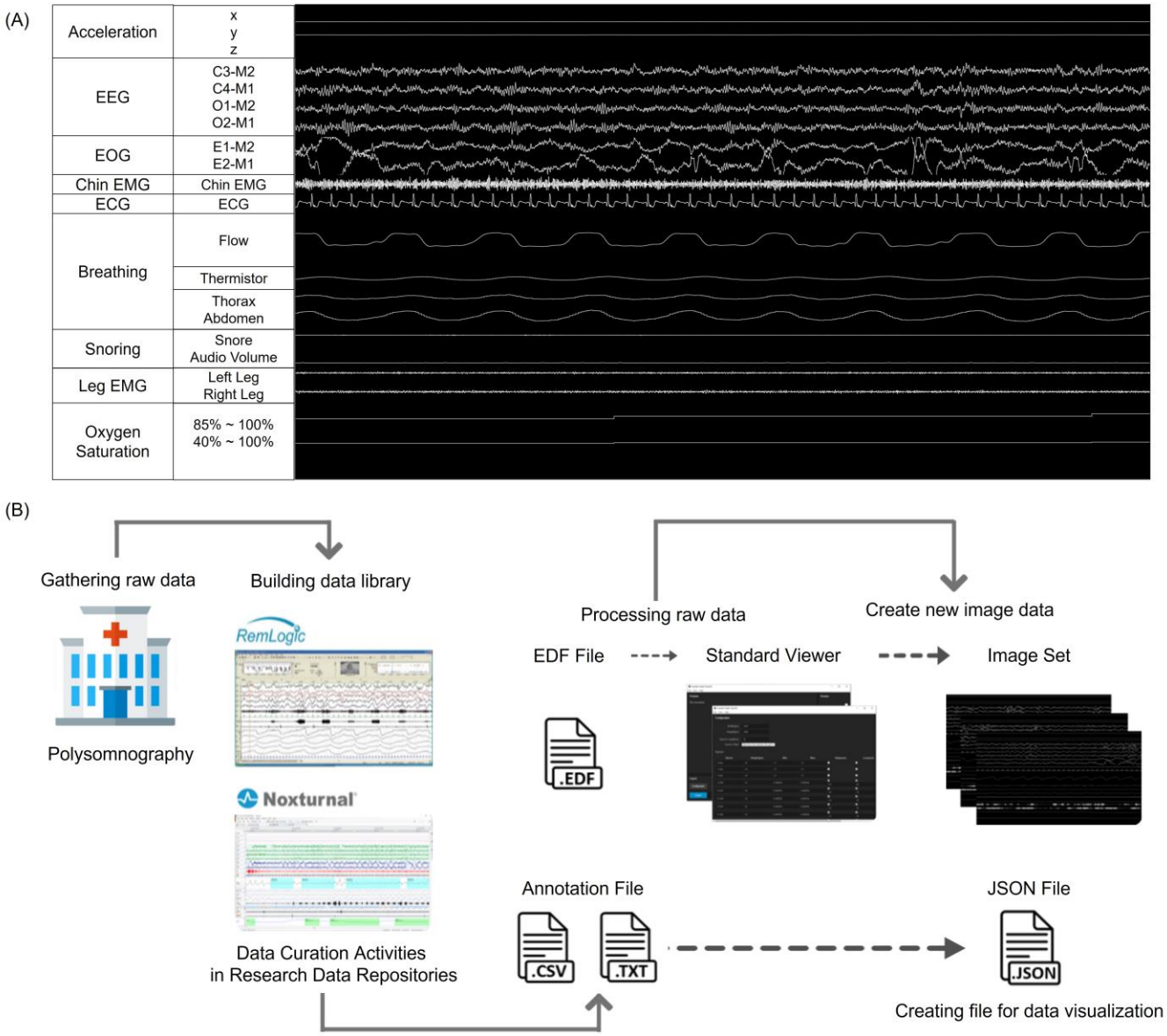


Figure 2

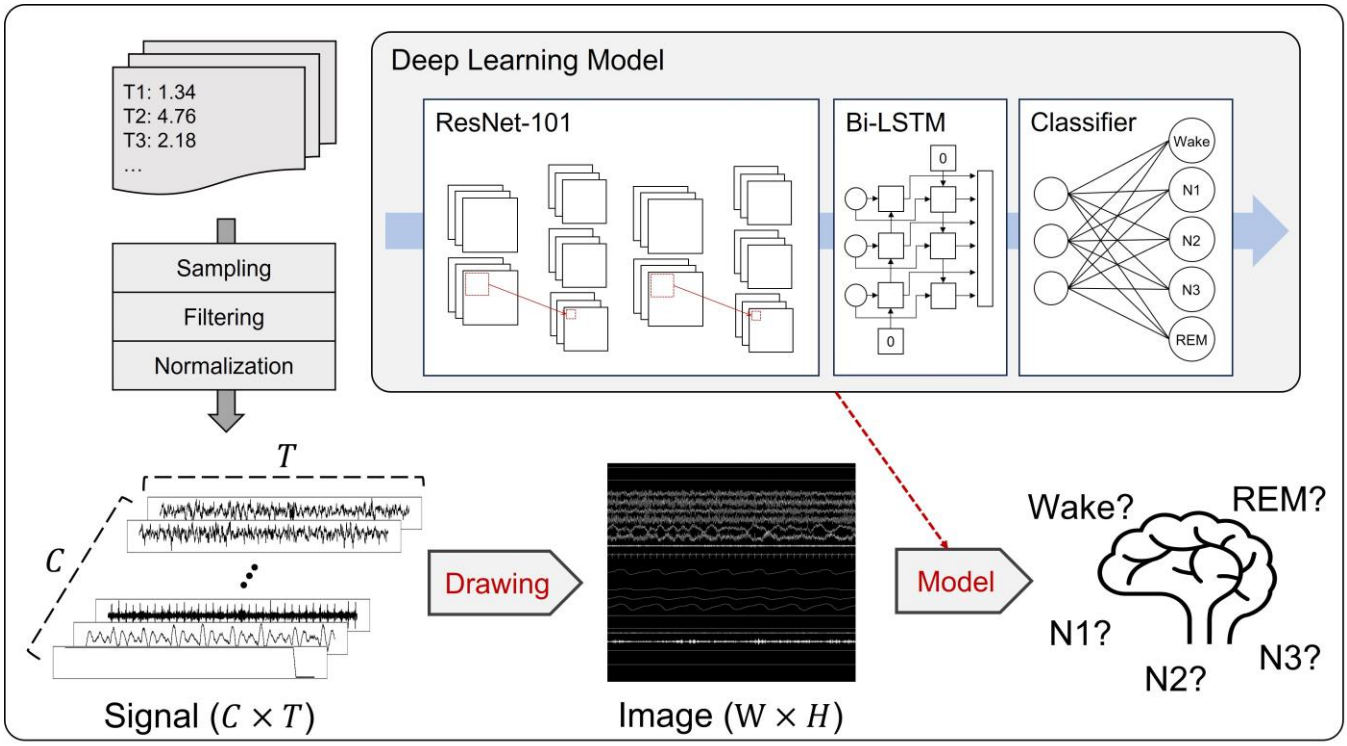


Figure 3

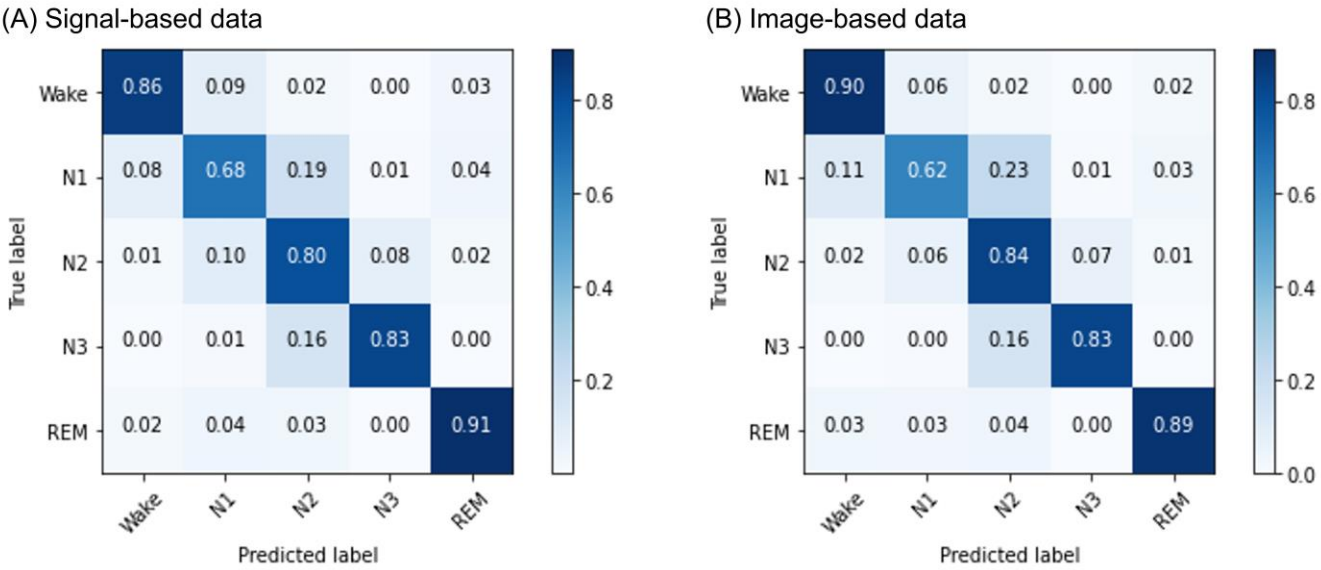


Figure 4

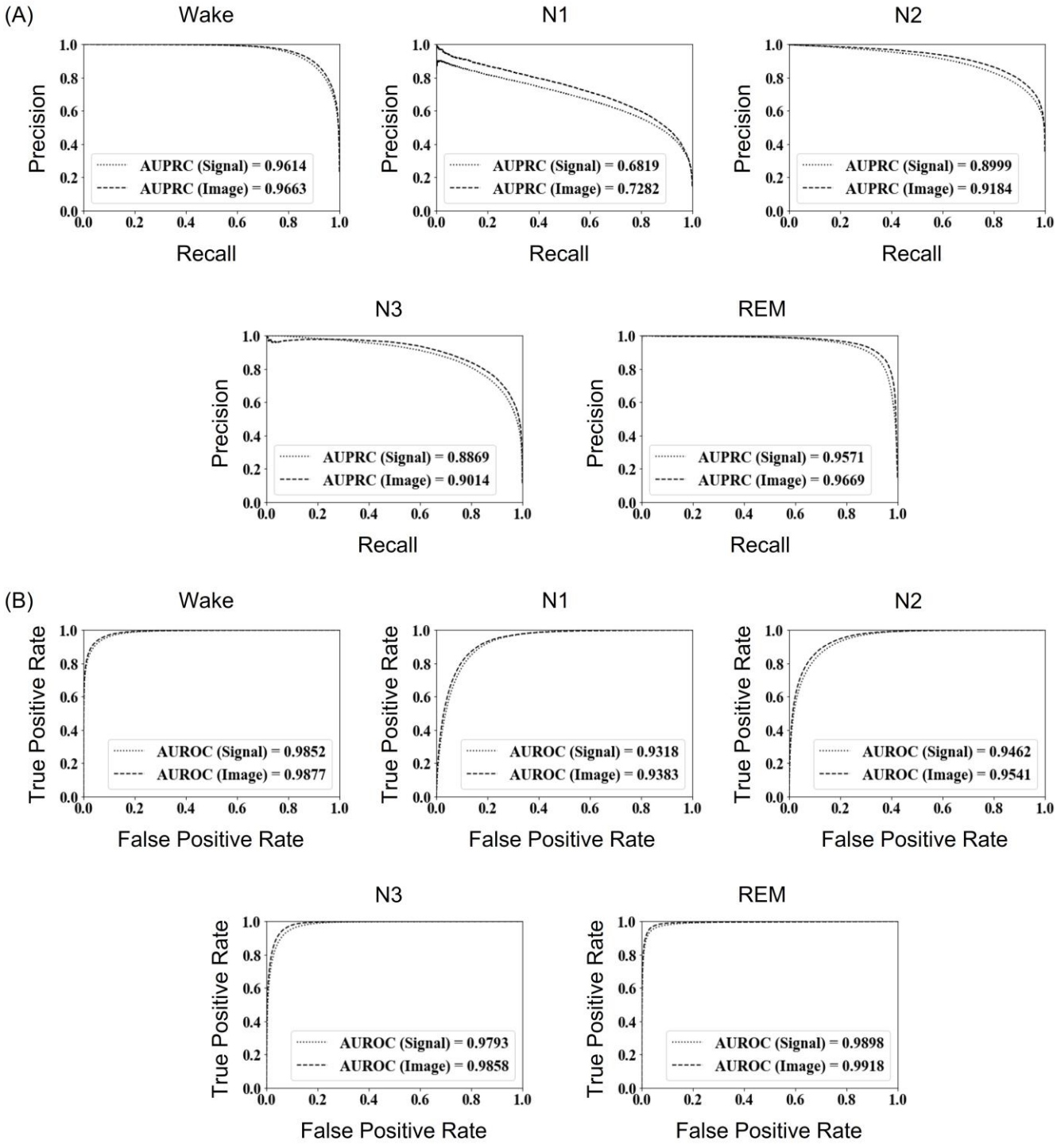


Figure 5

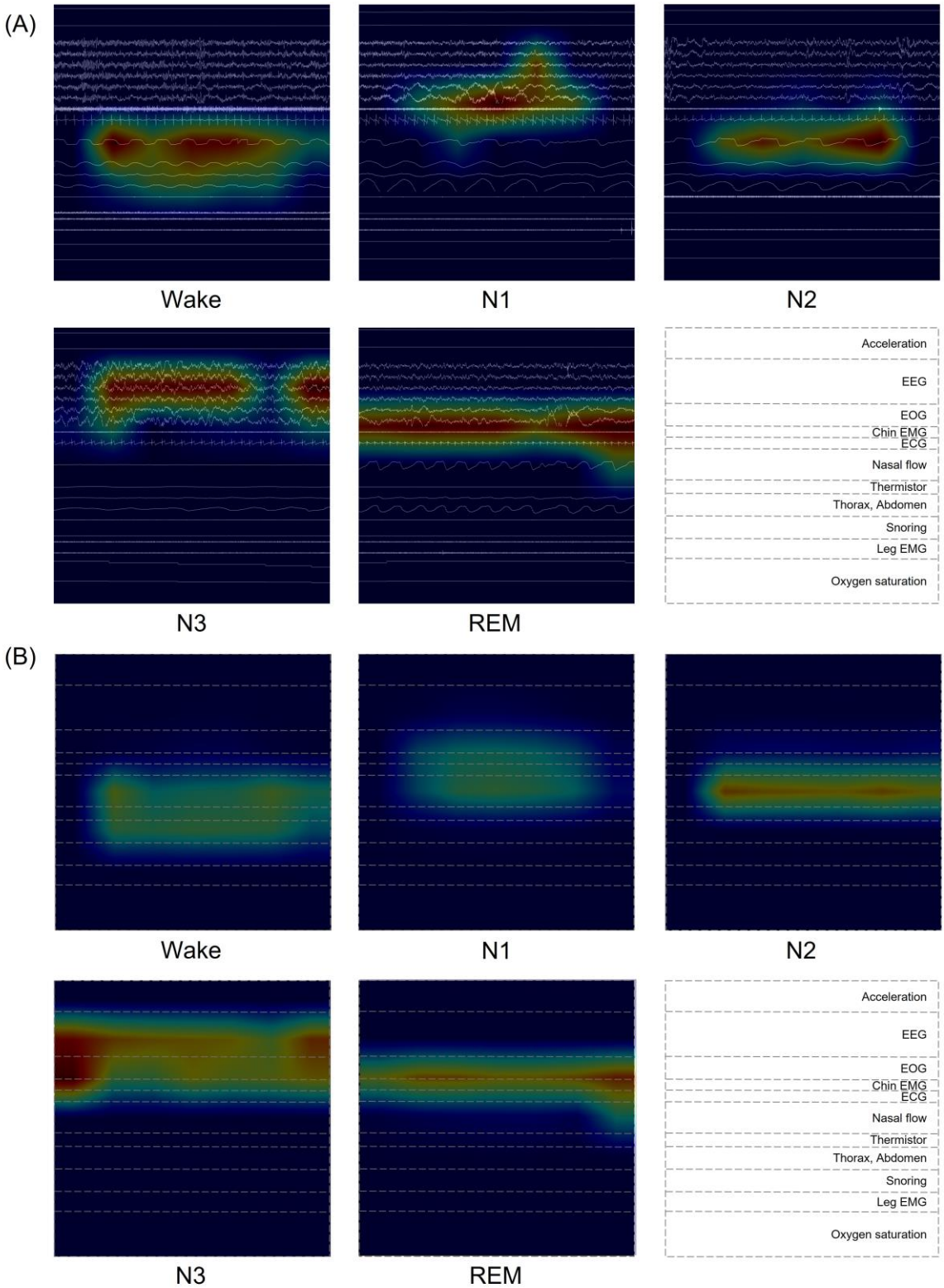


Figure 6

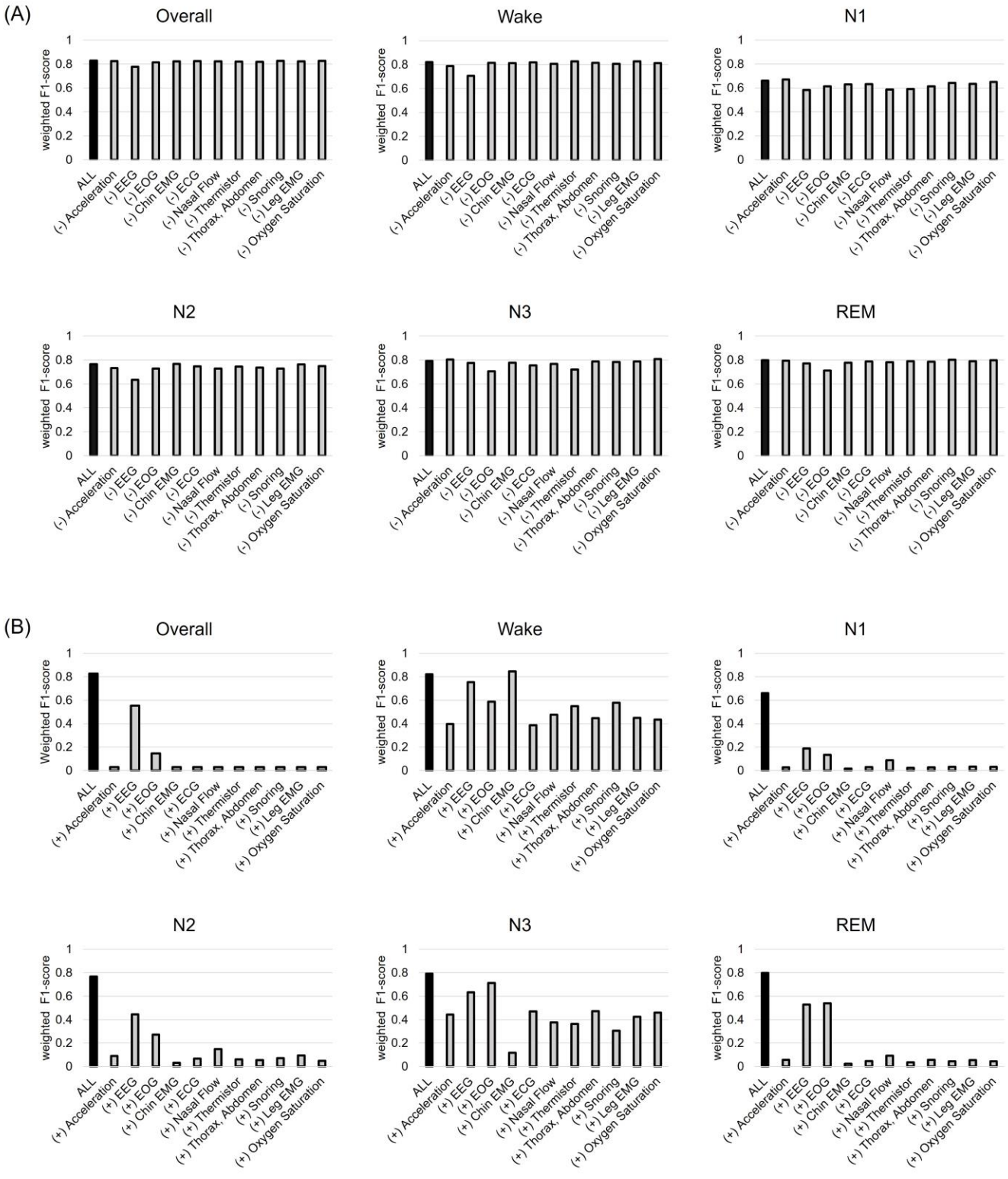
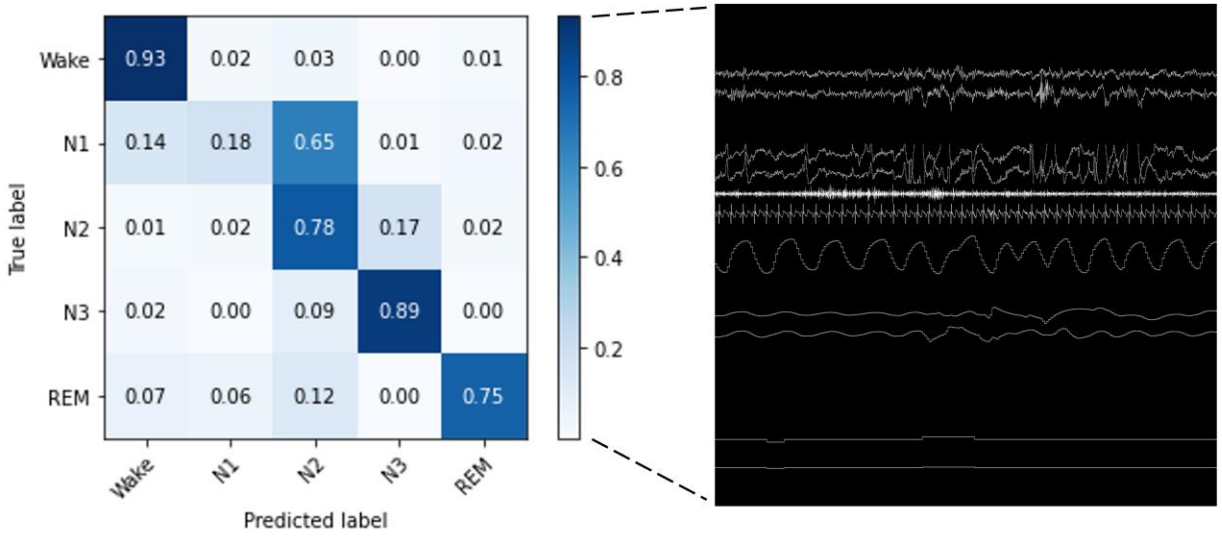
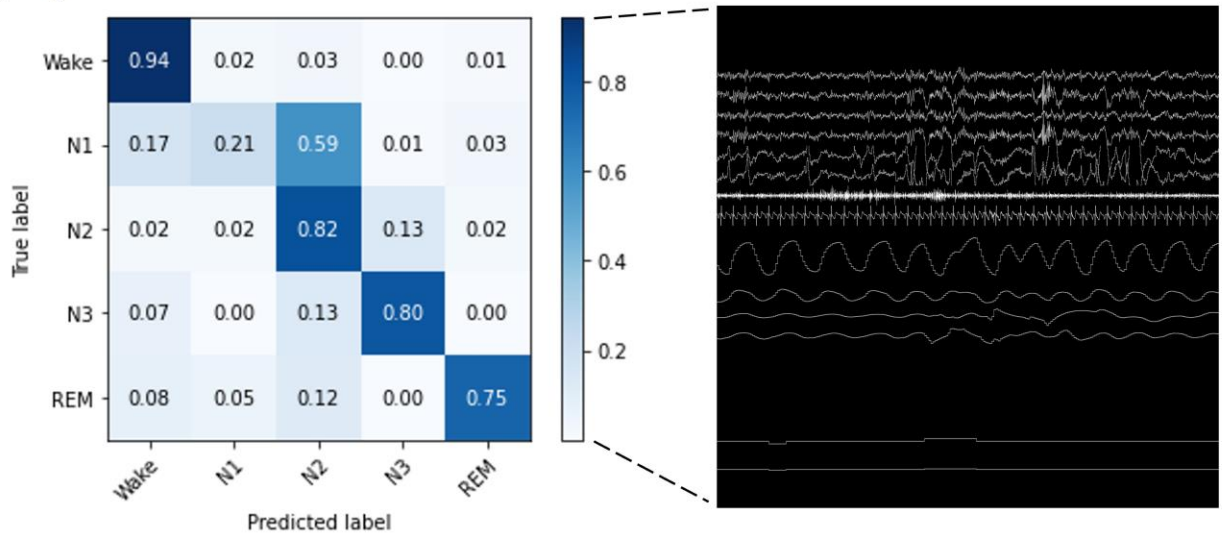


Figure 7

(A) Blanked



(B) Duplicated



(C)

FORMAT	Macro F1-Score	Micro F1-Score	Weighted F1-Score
Blanked	66.01%	79.32%	79.70%
Duplicated	69.24%	81.08%	81.70%