Patterns

Auto-annotating sleep stages based on polysomnographic data

Highlights

- Polysomnography enables accurate annotation of sleeping stages by machine learning
- Apnea/arousal can be more accurately detected by full polysomnography than EEG
- U-net achieved excellent performance in sequence-to-sequence prediction
- Our deep learning model achieves human-level accuracy in sleep status annotations

Authors

Hanrui Zhang, Xueqing Wang, Hongyang Li, Soham Mehendale, Yuanfang Guan

Correspondence

gyuanfan@umich.edu

In brief

In this study, we developed an automatic and fast sleeping stage and arousal/apnea detection tool, by adapting a U-net architecture with a convolutional neural network that is suitable for processing temporal information and makes sequence-to-sequence annotations. Our model is tested on different modalities and is consistently achieving excellent performance, which is comparable with human experts. Our tool provides an alternative to assist human experts in detecting pathological sleeping patterns in the study of clinical patients.

Zhang et al., 2022, Patterns *3*, 100371
January 14, 2022
https://doi.org/10.1016/j.patter.2021.100371

CellPress

Patterns

CellPress
OPEN ACCESS

Article

Auto-annotating sleep stages based on polysomnographic data

Hanrui Zhang,[1] Xueqing Wang,[1] Hongyang Li,[1] Soham Mehendale,[1] and Yuanfang Guan[1,2,3,*]
[1]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA
[2]Department of Internal Medicine, the University of Michigan Medical School, Ann Arbor, MI 48109, USA
[3]Lead contact
*Correspondence: gyuanfan@umich.edu
https://doi.org/10.1016/j.patter.2021.100371

THE BIGGER PICTURE Sleep quality is one of the top public health concerns. Disturbance during sleep will affect peoples' daily executive functions. In addition, some pathological sleeping conditions, such as arousal and apnea, are closely associated with severe health conditions such as cardiovascular diseases. Traditional sleeping surveillance requires laborious human effort while maintaining a limited reproducibility. In this study, we present a fast automatic sleep annotation deep learning model with excellent performances. Our model can annotate sleeping stages as well as sleeping arousal/apnea at the same time, which provides insight for clinical diagnosis of sleeping patients.

Proof-of-Concept: Data science output has been formulated, implemented, and tested for one domain/problem

SUMMARY

Sleep disorders affect the quality of life, and the clinical diagnosis of sleep disorders is a time-consuming and tedious process requiring recording and annotating polysomnographic records. In this work, we developed an auto-annotation algorithm based on polysomnographic records and a deep learning architecture that predicts sleep stages at the millisecond level. The model improves the efficiency of the polysomnographic record annotation process by automatically annotating each record within 3.8 s of computation time and with high accuracy. Disease-related sleep stages, such as arousal and apnea, can also be identified by this model, which further expands the physiological insights that the model can potentially provide. Finally, we explored the applicability of the model to data collected from a different modality to demonstrate the robustness of the model.

INTRODUCTION

Sleep disorders, which are prevalent in the general population, are growing threats to people's quality of life.[1] Besides their own negative impacts, some of the sleep disorders may be associated with other complex conditions, such as cardiovascular and metabolic disorders,[2] weakened immune system,[3] and neurologic diseases.[4] The primary dataset for identifying sleep disorders is whole-night polysomnography (PSG), which measures multiple physiology signals including EEG (brain waves), ECG (heart rhythm), EOG (eye movement), EMG (muscle movement), airflow, etc.[5] However, annotating the PSG records with sleep stages, which is crucial for disease diagnosis, is a tedious process. The annotation of an 8-h PSG record may require 2 h of repetitive work of a human expert.[6] Besides, because the sleep-stage scoring manuals, for example, the widely used American Academy of Sleep Medicine (AASM) scoring manual, have some ambiguities that require individual interpretation, the scoring results demonstrate a high variability among well-trained, experienced technologists.[7,8] Therefore, the development of computational approaches that could improve the speed and reliability of the sleep-scoring process is of great importance.

Despite a variety of machine learning and deep learning algorithms developed for sleep-stage scoring, auto-annotation of PSG records is still not widely accepted for clinical use.[6] Several open-ended questions are still challenging the field. The first one is that most automatic sleep-scoring algorithms are trained on the PSG records of healthy individuals, and therefore applying these algorithms to patients with sleep disorders often fails.[6] A possible reason for this is that patients with sleep disorders usually experience disease-related events such as arousal and apnea, which the models trained on healthy individuals are not exposed to. Most sleep-scoring models are only focused on the prediction of rapid eye movement (REM) and non-rapid eye

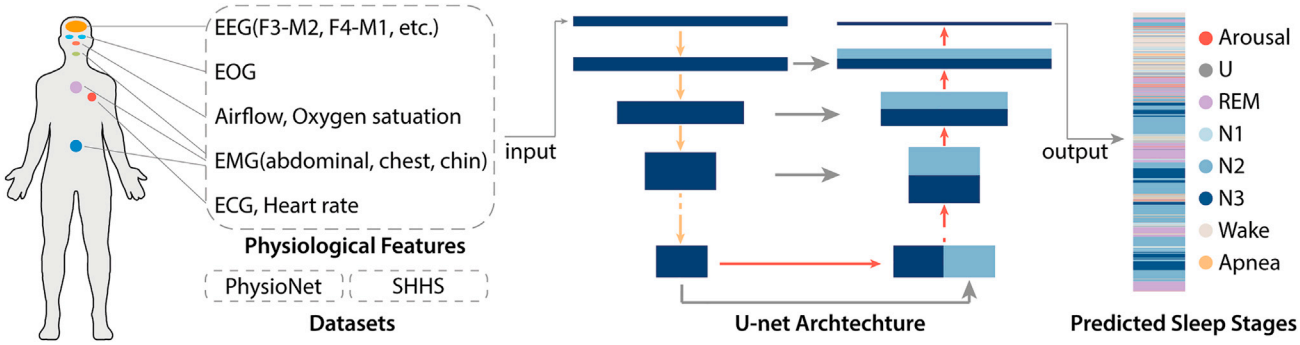


Figure 1. Overview of the experimental design

Diverse data collected from PhysioNet and SHHS were fed into the U-net, which segregates the time-series records into different sleep stages.

movement (NREM) stages. Ideally, an algorithm should also take arousal and apnea and all stages into consideration. Secondly, since the sleep-scoring manual is based on 30-s epochs, many algorithms are using 30-s epochs as input and make a prediction on these 30-s chunks.[9–12] However, it is not very accurate to view sleep stages as distinct entities, but they should be gradually transiting from one to another.[13]

In this work, we present a deep learning algorithm which auto-annotates the PSG records into seven categories, including five sleep stages (wake, N1, N2, N3, REM) and two pathological annotations (arousal, apnea), at the millisecond scale. We improved on the two questions mentioned above in two ways: firstly, we combined the prediction of sleep arousal and apnea with the prediction of basic sleep stages (wake, REM, NREM), which would be helpful to gain disease-related insights. The model was trained and tested separately on both PhysioNet dataset provided by the 2018 PhysioNet Challenge[14] and Sleep Heart Health Study visit 1 (SHHS-1) dataset provided by the National Sleep Research Resource (NSRR). Both datasets include a mix of healthy individuals and individuals with sleep disorders, making the resulting model applicable for the segmentation of disease-related records. Secondly, full-length sleep records are used as inputs and the U-net architecture is used to integrate information at different resolutions. The model achieves an annotation speed of 3.8 s for each record (on Nvidia Titan RTX) and a high accuracy (0.9826–0.8913 AUROC [area under the receiver operating characteristic curve]).

RESULTS

In this work, we developed a deep convolutional neural network model that can predict sleep stages based on overnight physiological measurements (Figure 1). The model was first developed and tested using the PhysioNet dataset (Table S1), and then validated using the SHHS-1 dataset (Table S2). To integrate both short-range and long-range information and produce predictions for each millisecond, our model adapted the U-net structure that takes the entire record as input. The model's performance was evaluated by the AUROC, the area under the precision-recall curve (AUPRC), precision, sensitivity, specificity, accuracy, and F1 score based on confusion matrix. Finally, we applied the models to predict the sleep patterns on the PhysioNet dataset and the SHHS-1 dataset and showed generalizability of the model.

Characteristics of the human-annotated PhysioNet dataset

The model was first developed using the datasets provided by the 2018 PhysioNet Challenge,[14] which is made up of 994 human-annotated polysomnography records. Each record consists of 13 physiology measurement channels, including 6 EEG channels collected using the International 10/20 system (F3-M2, F4-M1, C3-M2, C4-M1, O1-M2, O2-M1), 1 ECG channel, 1 EOG channel (left eye activity), 3 EMG channels (abdominal, chest, chin movement), 1 channel for the measurement of oxygen saturation (SaO_2), and 1 channel for the measurement of airflow.[14] All the signals are sampled at 200 Hz and are measured in microvolts.[14] The median record length is 7.7 h, with a standard deviation of 0.66 (Figure S1). The records were manually annotated by clinical staff according to the AASM manual for the scoring of sleep.[14] A total of seven types of sleep status annotations, including five sleep stages (wakefulness, NREM stages 1, 2, 3, REM) that are annotated by 30-s contiguous intervals, and two pathological symptoms that interrupt the sleep, arousal, and apnea.[14] NREM stage 2 (N2) and wakefulness exist with higher percentages while undefined and arousal have lower percentages (Figure S2). Detailed information about this dataset, including patient characteristics and annotation methods, can be found in supplementary materials (Tables S1 and S2).

Deep learning integrates information at different scales and enables localization

The model was adapted from the classic U-net architecture. Original U-net architecture was designed for 2D image segmentation.[15] We made two modifications: first, by alternating binary segmentation to multiple channels of output for different stages. Specifically, for each sleep stage, we have one channel representing it, with 1 representing that the time point falls into this particular sleep stage, and 0 otherwise. Secondly, we modified the 2D convolutional neural network to 1D convolutional neural network. Specifically, the convolution kernels are set up to 1D kernels (nX1), and the max pool layer is set up to pull along the longitudinal direction. The network consists of an encoder part and a decoder part. Each step in the encoder is made up of two convolutional layers and one downsampling layer, while each step in the decoder includes one upsampling layer and two deconvolutional layers. The high-resolution information from the encoding steps is duplicated and concatenated to the

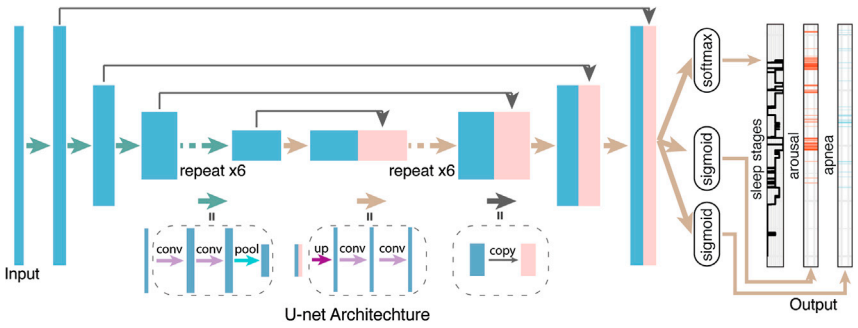


Figure 2. Model illustration

The model is made up of an encoder part that extracts information and a decoder part that extracts the information into sleep stages, including a total of 53 convolutional layers. Different activation functions are used for final output for sleeping stage, arousal, and apnea annotation.

corresponding decoding steps, allowing the precise localization of prediction results and capturing of the objects at different resolutions[15] (Figure 2). In addition, in contrast to previous models for sleep-stage annotation that typically use 30-s fragments as input, we used the entire record as input. In this way, both short-range and long-range information is considered, which improves the performance of the model.[16] A total of 11 steps in the encoder part and 11 steps in the decoder part are used, including a total of 47 convolutional layers, 11 maxpooling layers, and 11 upsampling layers. Padded convolutional layers are used to keep the output the same length as the input.

Data partition and augmentation

For this study, the 994 annotated data are divided into a training set with 795 records (80%) and a test set with 199 records (20%). Among the 795 records in the training set, 636 records (80%) are used for model training and 159 records (20%) are used for model validation. Training loss and validation loss are monitored to prevent overfitting. In addition, to simulate the random noise of the data, we randomly modified the input signal with a random factor between 0.95 and 1.05.[16]

Model performance and visualization

We developed the model using the 12 channels, excluding SaO_2, as the addition of this channel precludes convergence. To evaluate the model performance, we used 5-fold cross-validation by holding out 20% of the patients as the test set. We trained a total of five models and the mean outputs of these five models were used as the final output for each test record. AUROC, AUPRC, accuracy, sensitivity, specificity, precision, and F1 score are used as evaluation metrics, and each time point is used as one example in evaluation (Figure 3). The mean AUC scores for each stage range from 0.8913 (arousal) to 0.9826; the mean AUPRC scores for each stage range from 0.4121 (arousal) to 0.9183 (REM) (Figure 3A), as the values of AUPRC will largely depend on the fraction of each category in the entire time course. Figure 3B shows the confusion matrix of apnea and arousal prediction, as well as the prediction for five sleeping stages, where the sensitivity of sleeping stage prediction ranges from 54.37% (N1) to 92.13% (REM). For the model's speed, the average time taken to make predictions and evaluate one record was only 3.8 s (Figure S3), which is a huge improvement on the speed of human annotation. Instead of generating an overall prediction label for a 30-s chunk as in most of the conventional prediction models, a prediction label is generated for each data point, which increases the resolution of prediction results. To visualize the prediction results, we displayed a prediction example on an example record tr05-1452 (Figure 4). Figure 4A shows the original 12-channel polysomnography signal of this record. Figures 4B and 4C show the gold standard and prediction of arousal/apnea and the five sleeping stages, respectively. From these two figures, we can see that the sleep stages are dynamically changing between light and deep sleep stages during the course of sleep, forming several cycles. Figure 4D shows the full evaluation matrices of the prediction performance on this record. These visualizations would be helpful for clinicians to easily understand the model output and assess the model performance on a single record.

Model validation on SHHS data

To validate the model, we used the SHHS-1 dataset from NSRR. Of note, the channels of this study cannot be directly mapped to the PhysioNet study, and for the couple that do match, they were collated in different modalities.

This dataset provided PSG records of 5,793 participants (healthy or with various kinds of sleep problems). There are 16 channels including 2 EEG channels (C4-A1, C3-A2), 1 ECG channel, 2 EOG channels (left and right eye activities), 2 EMG channels (lower chin movement), 2 channels for airflow, and 1 channel for each of the following items: oxygen saturation, heart rate, position, light, sound, and oximetry. All channels were sampled at a frequency of 125 Hz. The two EEG-driven grounds were placed on C3 and C4, and paired with nodes on A1 and A2. The ECG nodes were placed 3–5 cm below the middle of the right and left collar bones, in spaces between rib bones; the two EOG nodes were placed 1 cm out and 1 cm down from the outer corner from the right and left eyes, on the bony ridge, and two EMG nodes were placed on the lower chin, separated by 1 cm.[17] The annotations of the records are summarized into six stages including wake, sleep stages 1, 2, 3, 4, REM, where stages 3 and 4 are combined as deep sleep stage;[17] arousal and apnea during sleep are also annotated in parallel alongside the sleep stages.

We randomly selected 1,000 records from the 6,441 records in the SHHS-1 dataset, which is comparable with the 994 records in the PhysioNet dataset. The mean length of these 1,000 records is 8.4 h, with a standard deviation of 0.61 (Figure S4). Compared with the PhysioNet dataset, the SHHS-1 dataset has a lower average percentage of "N1" stage, while having a higher average percentage of "apnea" and "wake" stages (Figure S5). The 1,000 records are randomly divided into a training set with 800 records (80%) and a test set with 200 records (20%). The eight channels that overlap between the SHHS-1 dataset and Fox dataset were used to develop the model, including "SaO_2," "H.R.," "ECG," "THOR RES," "ABDO RES," "POSITION," "LIGHT," and "OX stat." A total of five

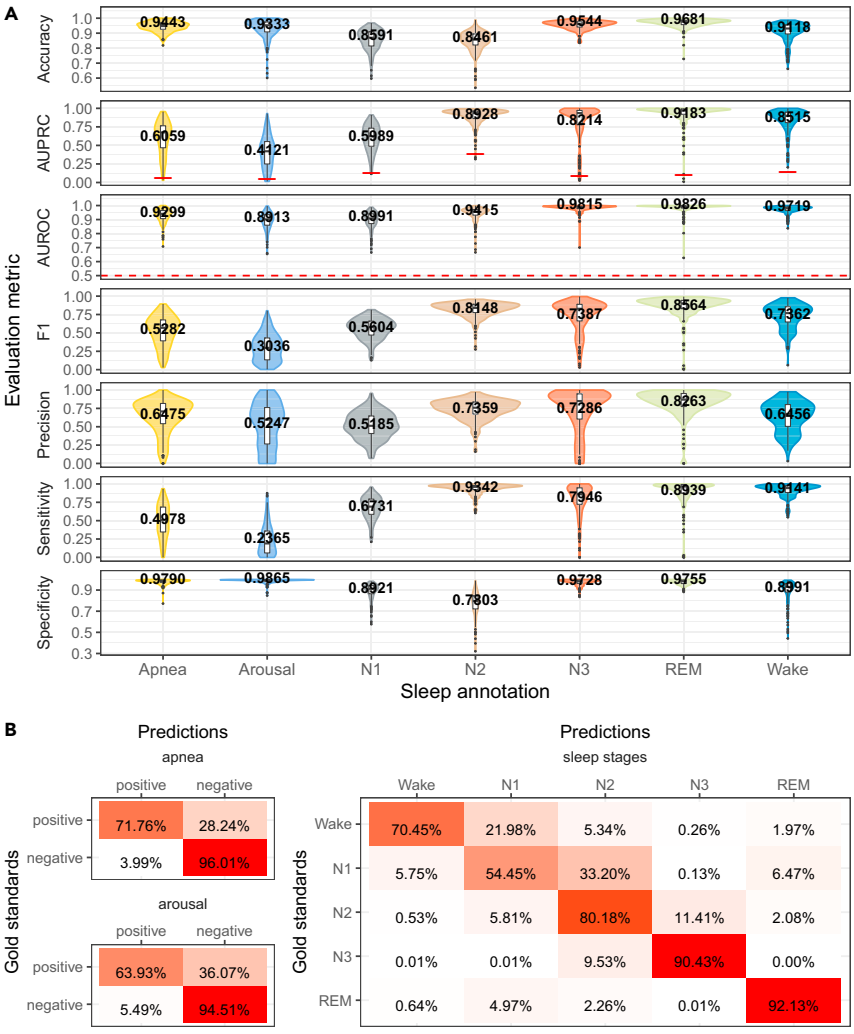


Figure 3. Model performance of each sleep annotation on the PhysioNet dataset (199 test samples in total)

(A) Model performances evaluated by: accuracy, AUPRC, AUROC, F1, precision, sensitivity, and specificity. The average performances of all samples in the test sets for each label are marked in the figures. Baselines of AUROC and AUPRC are marked by red dashed lines. The violin plot indicates the model's performance on the test set.

(B) Confusion matrix of each sleep annotation of model predictions of all samples in the test set for PhysioNet dataset.

We noticed for PhysioNet dataset, when using the 6 EEG channels as input, the performance of the 5 sleeping stages (N1, N2, N3, REM, and wake) is on a par with using all 12 channels (0.9808 (REM)–0.9069 (N1) AUROC), while the arousal and apnea prediction performance decreased noticeably (0.8327 and 0.8541 AUROC). This is probably because the sleeping stages are denoted based on EEG signals solely, while apnea and arousal status also involves others, such as body movement. For the SHHS-1 dataset, using the two EEG signals only did not achieve satisfactory performances (about the random baseline), probably due to a lack of information.

DISCUSSION

In this work, we developed a sleep-stage auto-annotation algorithm for PSG records, which can annotate the record into five sleep stages (wake, stage 1, 2, 3, REM) and two pathological stages (arousal and apnea) at a speed of 3.8 s per record and with high accuracy (AUC = 0.9826–0.8913). The use of U-net architecture improved the performance of the model, and including the disease-related stages (arousal and apnea) provides more insights than basic models.

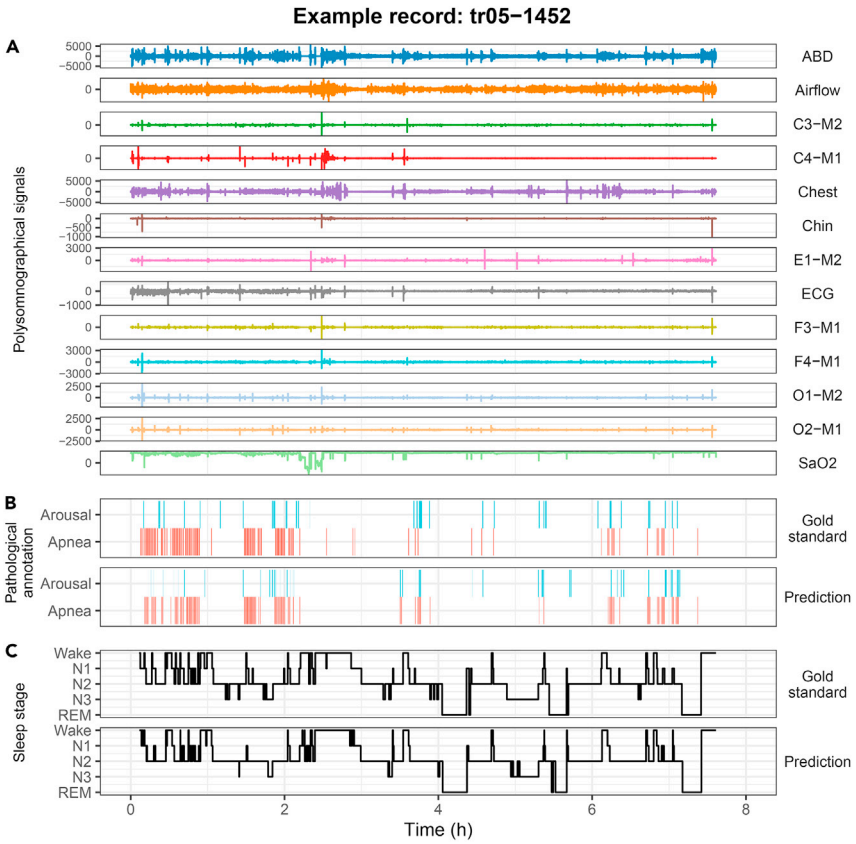
Previous sleeping stage annotation models have always employed recurrent neural networks, long short-term memory networks (LSTM), and convolution neural networks (CNN)[12,18,21–26] to process continuous temporal signals. Our model in this study, which is U-net architecture from CNN, achieved the state of the art prediction performances compared with previous works, including XSleepNet,[21] RCNN,[25] and the CNN/LSTM ensemble model[27] on the same SHHS or PhysioNet dataset (Table S3), while being able to generate multiple outputs at the same time from a single network, including five sleep-stage prediction and annotations for apnea and arousal. U-net was first invented for and achieved excellent performance in biomedical image segmentation tasks.[15] In U-net, the copy and crop of previous layers during upsampling helps retain the original information from the input image, therefore improving the precision of localization when mapping the segments back to the original input.

A major roadblock that impedes the clinical use of sleep-stage auto-annotation models is that datasets collected by different

models were trained, and the average result of the five models was considered as the final result. The performance of the model on annotating the sleep stages of the SHHS-1 test set is similar to the model developed using 12-channel PhysioNet data, with an AUROC of 0.8584 (N1)–0.9721 (REM) and AUPRC of 0.2218 (N1)–0.9223 (wake) (Figure S6). To further explore the generosity of our model, we also tried 1:1 training/testing set splits for model training and validation for SHHS-1. The model maintained robust prediction performances with decreased training sets and increased unseen testing sets, achieving 0.9651 (wake)–0.8200(N1) AUROC, 0.9102 (wake)–0.1610 (N1) AUPRC, and 95.68% (arousal)–72.6% (N2) accuracy (Figure S8).

Comparison of a full polysomnography model with EEG as input

While our sleeping stage annotation model is based on the full polysomnography, there are also many previous sleeping stage annotation studies based on EEG signals only.[18–20] Therefore, we also carried out the sleep-stage prediction task with our model input with EEG signals only. The performance of using EEG signals only for PhysioNet and SHHS dataset prediction is shown in Figures S9 and S10.



D

Label	AUROC	AUPRC	Sensitivity	Specificity	Precision	Accuracy	F1
Arousal	0.9310	0.4867	0.3586	0.9760	0.4973	0.9377	0.4167
Apnea	0.9565	0.7292	0.5533	0.9814	0.7401	0.9441	0.6332
Wake	0.9891	0.9481	0.9603	0.9544	0.8094	0.9554	0.8784
N1	0.9588	0.7650	0.7857	0.9381	0.6096	0.9215	0.6866
N2	0.9760	0.9774	0.9910	0.6989	0.7757	0.8486	0.8702
N3	0.9858	0.8799	0.8094	0.9780	0.7795	0.9633	0.7942
REM	0.9979	0.9867	0.9325	0.9971	0.9750	0.9901	0.9533

Figure 4. Example of prediction on one sample (tr05-1452) in PhysioNet dataset

(A) Original signal with 13 channels as input.

(B) Gold standard versus prediction for the "arousal" and "apnea" sleep status annotations.

(C) Hypnography of the gold standard versus prediction for five sleeping stages annotations.

(D) The prediction performances by seven metrics (AUROC, AUPRC, sensitivity, specificity, precision, accuracy, and F1) on eight different labels, respectively.

devices sometimes significantly differ from each other, therefore the model performs less satisfying when the training data and test data come from different sources. Nevertheless, we found that the modification of the U-net structure is applicable separately to data collected based on drastically different modalities.

We noticed a major limitation of our multi-class sleep-stage annotation model is the severely imbalanced data distribution in sleep stages, represented by an overall lower prediction accuracy of N1 stage in both the PhysioNet and SHHS-1 datasets. One reason could be that there is a tradeoff between achieving the overall lowest validation loss versus the lowest loss on each class, respectively. Therefore the classification accuracy for N1 could be compromised by other more dominant sleeping stages during training, such as wake, REM, and especially "N2," which occupied about half of the sleeping cycle (over 40% for both PhysioNet and SHHS-1). In addition, we observed that a large portion of N1 were misclassified as N2 from the confusion matrix in all classification experiments, leading to a lower sensitivity of N1 detection (Figures 3B, S6B, and S8B). A lower detection sensitivity of N1 was also consistently observed in other previous machine-leaning sleep-stage annotation studies.[22,24,25] During human sleep, the N1 stage is usually very short (1–5 min) and the transition between N1 to N2 is fast,[5] which probably contributes to the increasing difficulty to capture the exact boundary between N1 and N2, especially for human experts. As previously reported in another study, human experts achieved only 46% precision and 48% recall in N1 annotation,[22] while our model achieved 51.18% precision and 67.31% recall. Therefore, the "gold standard," or human annotation for N1 stage may not be truly the ground truth, bringing a challenge to predicting a perfect score. As a matter of fact, contrary to the N1 stage, the REM stage of sleep, while also only taking up one-eighth of total sleep cycle, as characterized by increasing brain activity and eye movement, achieved an astoundingly excellent accuracy by our model (0.9826 average AUC), probably due to the much more accurate and consistent ground truth annotation by human experts.

Finally, since the technical setup of polysomnography is complex, sleep monitoring can only be carried out in a lab setting. To facilitate the setup of convenient home monitoring systems, essential signals for sleep staging should be identified. Future work should explore the ways to improve model performance in the situations where the test data and training data are from different sources, and the minimum signals required for accurate sleep-stage scoring.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Yuanfang Guan (gyuanfan@umich.edu).

Materials availability

This paper analyzes existing, publicly available data. The PhysioNet and SHHS datasets can be accessed from the 2018 PhysioNet Challenge website (https://physionet.org/physiobank/database/challenge/2018/) and the Sleep Heart Health Study website (https://sleepdata.org/datasets/shhs).

Data and code availability

All original code has been deposited at Github under https://github.com/GuanLab/Sleep-Stage-Auto-annotation and is publicly available as of the date of publication.

Overview of datasets and models

Two datasets were used in this paper: the PhysioNet dataset from the 2018 PhysioNet Challenge[14] and the SHHS-1 dataset from NSRR. For the PhysioNet dataset, 41.8% of the patients visited the sleep laboratory for diagnostic reasons, and 58.2% visited for CPAP, according to the data descriptions from the data provider. The SHHS study includes 27.2% participants from the Atherosclerosis Risk in Communities Study (ARIC), 21.0% from the Cardiovascular Health Study (CHS), 15.5% from the Framingham Heart Study (FHS), 9.3% participants from the Strong Heart Study (SHS), 15.5% from the New York Hypertension Cohorts, and 14.0% from the Tucson Epidemiologic Study of Airways Obstructive Diseases and the Health and Environment Study. The datasets were used in two training processes and four prediction processes. The PhysioNet dataset and the SHHS-1 dataset were used to train two sets of deep learning models separately. To obtain more robust results, five models were trained for each set, and all the results generated were the average result of the five models. Each of the two sets of models were used to predict the sleep stages of two datasets separately.

Data partition

The 994 records from the PhysioNet dataset and the 1,000 randomly selected records from the SHHS-1 dataset were randomly divided into two parts: 80% of the records were used as the training set and 20% of the records were used as the test set. Then the training set was further divided into a second training set (80%) and a validation set (20%). The validation set was used to monitor the model training process and prevent overfitting. In each training epoch, the model was first trained using the records in the second training set, and then the records in the validation set were provided to the model to calculate a validation loss without changing model weights. We stopped training the model before the validation loss rebounded to prevent overfitting.

Quantile normalization

To account for the technical differences when generating each record, we normalized the records using quantile normalization.[28,29] First, a reference record was generated by resizing every record in a dataset into the same length, sorting each of them, and then averaging the sorted records. Then the highest value in the target record was replaced by the highest value in the reference record, the second highest value was replaced by the second highest record, and so on. Finally, the target record would have the same values as the reference record, while the sequence of the values in the target record was preserved. Two reference records were generated, one using the PhysioNet dataset and the other one using the SHHS-1 dataset. Before being fed into the models trained by either of the datasets, all the input records were quantile normalized for each channel using the corresponding reference record.

Loss function

The loss function used in this model is weighted cross-entropy loss. To understand this loss function, we first introduce the cross-entropy loss function, which is defined as:

$$H(y,\hat{y}) = \sum_{i=1}^{N} [-y_i \times \log\hat{y_i} - (1-y_i) \times \log(1-)\hat{y_i}],$$

where y_i refers to the gold standard label at time point *i*, and $\hat{y_i}$ refers to the predicted label at time point *i*. Note that this equation is used to calculate loss for a single stage, for example, whether the sleep stage at a time point belongs to stage N1 (label = 1) or not (label = 0). To develop a loss function that represents the model's loss on all sleep annotations, we assigned different weights for calculating the loss for the five sleeping stages, arousal and apnea (0.18:0.6:0.22), to synchronize the converging process for all three tasks. Arousal was assigned a larger weight because it converged slower than the other two.

Model training

To fit the capacity of the GPU memory, we average-pooled the length of input PhysioNet records to one-eighth of the original resolution and the SHHS-1 records to one-fifth of the original resolution. Specifically, for PhysioNet data, we used the average value of every eight points as a new data point and, for SHHS-1 data, we used the average value of every five points as a new data point. The labels were one-hot encoded. Adam optimizer was used in the training process, with a learning rate of 0.0001. We monitored the training loss and validation loss, and stopped training the model before the validation loss rebounded to prevent overfitting. The model has multiple channels of output representing different stages. The models trained by the PhysioNet dataset were trained for 40 epochs and the models trained by the SHHS-1 dataset were trained for 50 epochs.

Model evaluation

The performance of the model was measured by the AUROC and the AUPRC. In this work, the ROC curve is generated by plotting the true-positive rate (TPR) against the false-positive rate (FPR) at 1,000 thresholds, and the precision-recall curve is created by plotting the precision against the recall at 1,000 thresholds. TPR (recall), FPR, and precision are defined as below:

$$TPR(recall) = True\ Positive/(True\ Positive + False\ Negative)$$
$$FPR = False\ Positive/(False\ Positive + True\ Negative)$$
$$Precision = True\ Positive/(True\ Positive + False\ Positive).$$

The AUC was estimated by summing up the area of each rectangle bin. AUC and AUPRC scores were calculated for each channel in each test record. The average AUC and AUPRC score for each channel was calculated by averaging scores from all the test records.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2021.100371.

ACKNOWLEDGMENTS

This study was supported by NIH Project 1R35-GM133346-01: Machine Learning for Drug Response Prediction; by Michael J. Fox Foundation Project no. 17373: Interpretation of Accelerometer Data with deep learning to Y.G.; and by American Heart Association and Amazon Web Services Data Grant Portfolio 3.0: Artificial Intelligence and Machine Learning Training Grants Award 19AMTG34850176 to H.L. We thank the GPU donation from Nvidia and the AWS donation from Amazon.

AUTHOR CONTRIBUTIONS

Conceptualization, Y.G.; methodology, Y.G.; writing – original draft, X.W., writing – review & editing, H.Z. and Y.G.; formal analysis and investigation, X.W. and H.Z.; visualization, X.W. and H.Z.; literature and demographic survey, S.M.; supervision, Y.G. and H.L.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: June 21, 2021
Revised: July 15, 2021
Accepted: September 28, 2021
Published: October 28, 2021

REFERENCES

1. Ohayon, M.M. (2011). Epidemiological overview of sleep disorders in the general population. Sleep Med. Res. *2*, 1–9.
2. Tobaldini, E., Costantino, G., Solbiati, M., Cogliati, C., Kara, T., Nobili, L., and Montano, N. (2017). Sleep, sleep deprivation, autonomic nervous system and cardiovascular diseases. Neurosci. Biobehav. Rev. *74*, 321–329.
3. Besedovsky, L., Lange, T., and Haack, M. (2019). The sleep-immune crosstalk in health and disease. Physiol. Rev. *99*, 1325–1380.
4. Barone, D.A., and Chokroverty, S. (2017). Neurologic diseases and sleep. Sleep Med. Clin. *12*, 73–85.
5. Patel, A.K., Reddy, V., and Araujo, J.F. (2020). Physiology, sleep stages. In StatPearls (StatPearls Publishing).

6. Fiorillo, L., Puiatti, A., Papandrea, M., Ratti, P.-L., Favaro, P., Roth, C., Bargiotas, P., Bassetti, C.L., and Faraci, F.D. (2019). Automated sleep scoring: a review of the latest approaches. Sleep Med. Rev. *48*, 101204.

7. Younes, M., Kuna, S.T., Pack, A.I., Walsh, J.K., Kushida, C.A., Staley, B., and Pien, G.W. (2018). Reliability of the American Academy of Sleep Medicine rules for assessing sleep depth in clinical practice. J. Clin. Sleep Med. *14*, 205–213.

8. Danker-Hopfe, H., Kunz, D., Gruber, G., Klösch, G., Lorenzo, J.L., Himanen, S.-L., Kemp, B., Penzel, T., Röschke, J., Dorn, H., et al. (2004). Interrater reliability between scorers from eight European sleep laboratories in subjects with different sleep disorders. J. Sleep Res. *13*, 63–69.

9. Correa, A.G., Laciar, E., Patiño, H.D., and Valentinuzzi, M.E. (2008). An automatic sleep-stage classifier using electroencephalographic signals. Int. J. Med. Sci. *1*, 13–21.

10. 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. https://doi.org/10.1016/j.neucom.2012.11.003.

11. Alickovic, E., and Subasi, A. (2018). Ensemble SVM method for automatic sleep stage classification. IEEE Trans. Instrum. Meas. *6c7*, 1258–1265.

12. 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.

13. Malhotra, R.K., and Avidan, A.Y. (2013). Introduction to sleep stage scoring. Atlas Sleep Med. Expert Consult. Online Print, 77.

14. Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P.C., Mark, R., et al. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation [Online] *101*, e215–e220.

15. Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015 (Springer International Publishing), pp. 234–241.

16. Li, H., and Guan, Y. (2021). DeepSleep convolutional neural network allows accurate and fast detection of sleep arousal. Commun. Biol. *4*, 18.

17. Zhang, G.-Q., Cui, L., Mueller, R., Tao, S., Kim, M., Rueschman, M., Mariani, S., Mobley, D., and Redline, S. (2018). The National Sleep Research Resource: towards a sleep data commons. J Am Med Inform Assoc *25*, 1351–1358.

18. 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.

19. Aboalayon, K., Faezipour, M., Almuhammadi, W., and Moslehpour, S. (2016). Sleep stage classification using EEG signal analysis: a comprehensive survey and new investigation. Entropy *18*, 272. https://doi.org/10.3390/e18090272.

20. Satapathy, S.K., Ravisankar, M., and Logannathan, D. (2020). Automated sleep stage analysis and classification based on different age specified subjects from a dual-channel of EEG signal. In 2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT). https://doi.org/10.1109/conecct50063.2020.9198335.

21. Phan, H., Chen, O.Y., Tran, M.C., Koch, P., Mertins, A., and De Vos, M. (2021). XSleepNet: multi-view sequential model for automatic sleep staging. IEEE Trans. Pattern Anal. Mach. Intell. https://doi.org/10.1109/TPAMI.2021.3070057.

22. Sokolovsky, M., Guerrero, F., Paisarnsrisomsuk, S., Ruiz, C., and Alvarez, S.A. (2020). Deep learning for automated feature discovery and classification of sleep stages. Ieee/acm Trans. Comput. Biol. Bioinform. *17*, 1835–1845.

23. 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 Digit Med. *3*, 106.

24. Yildirim, O., Baloglu, U.B., and Acharya, U.R. (2019). A deep learning model for automated sleep stages classification using PSG signals. Int. J. Environ. Res. Public Health *16*, 599. https://doi.org/10.3390/ijerph16040599.

25. 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.

26. Warrick, P., and Homsi, M.N. (2018). Sleep arousal detection from polysomnography using the scattering transform and recurrent neural networks. Computing in Cardiology *45*. https://doi.org/10.22489/cinc.2018.368.

27. Warrick, P.A., and Nabhan Homsi, M. (2018). Ensembling convolutional and long short-term memory networks for electrocardiogram arrhythmia detection. Physiol. Meas. *39*, 114002.

28. Bolstad, B.M., Irizarry, R.A., Astrand, M., and Speed, T.P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics *19*, 185–193.

29. Li, H., and Guan, Y. (2021). Fast decoding cell type-specific transcription factor binding landscape at single-nucleotide resolution. Genome Res. *31*, 721–731.