SLEEPJ, 2019, 1–10

doi: 10.1093/sleep/zsz159
Advance Access Publication Date: 9 July 2019
Original Article

Original Article

Automated sleep stage scoring of the Sleep Heart Health Study using deep neural networks

Linda Zhang[1], Daniel Fabbri[1], Raghu Upender[2] and David Kent[3,*]

[1]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, [2]Department of Neurology, Sleep Disorders Division, Vanderbilt University School of Medicine, Nashville, TN and [3]Department of Otolaryngology, Vanderbilt University Medical Center, Nashville, TN.

*Corresponding author. David Kent, Department of Otolaryngology, Vanderbilt University Medical Center, 1215 21st Avenue South, Suite 7209, Nashville, TN 37232. Email: david.kent@vumc.org.

Abstract

Study Objectives: Polysomnography (PSG) scoring is labor intensive and suffers from variability in inter- and intra-rater reliability. Automated PSG scoring has the potential to reduce the human labor costs and the variability inherent to this task. Deep learning is a form of machine learning that uses neural networks to recognize data patterns by inspecting many examples rather than by following explicit programming.

Methods: A sleep staging classifier trained using deep learning methods scored PSG data from the Sleep Heart Health Study (SHHS). The training set was composed of 42 560 hours of PSG data from 5213 patients. To capture higher-order data, spectrograms were generated from electroencephalography, electrooculography, and electromyography data and then passed to the neural network. A holdout set of 580 PSGs not included in the training set was used to assess model accuracy and discrimination via weighted F1-score, per-stage accuracy, and Cohen's kappa (K).

Results: The optimal neural network model was composed of spectrograms in the input layer feeding into convolutional neural network layers and a long short-term memory layer to achieve a weighted F1-score of 0.87 and K = 0.82.

Conclusions: The deep learning sleep stage classifier demonstrates excellent accuracy and agreement with expert sleep stage scoring, outperforming human agreement on sleep staging. It achieves comparable or better F1-scores, accuracy, and Cohen's kappa compared to literature for automated sleep stage scoring of PSG epochs. Accurate automated scoring of other PSG events may eventually allow for fully automated PSG scoring.

Statement of Significance

Sleep staging is an important part of evaluating overnight polysomnograms. Sleep stages are scored by technicians and physicians based on visual examination of neurophysiologic signal patterns. This process is labor intensive and suffers from variability between scorers. In this study, large amounts of publicly available polysomnography data were used to train a sleep staging classifier. Sleep staging classification by the model achieved better agreement than human agreement in literature. Generalizability of the model to other unseen datasets from different public projects is also demonstrated.

Key words: polysomnography; sleep staging; deep learning; machine learning

Submitted: 21 September, 2018; **Revised:** 19 May, 2019

Introduction

Overnight polysomnography (PSG) is central to the diagnosis and management of many sleep disorders. The clinical standard for PSG sleep staging requires visual inspection of the data by trained sleep technicians and physicians. Staging historically followed the Rechtschaffen and Kales (RK) criteria until the American Academy of Sleep Medicine (AASM) published updated criteria in 2007 [1, 2]. The AASM rules divide sleep into five stages: Wake, Non-Rapid Eye Movement stages 1, 2, and 3 (N1, N2, and N3), and Rapid Eye Movement (REM). PSG scoring is a labor-intensive process that requires up to 2 hours for a sleep technologist to complete [3]. Inter-rater and intra-rater reliability of PSG staging and event scoring is also known to suffer from considerable variability [4–13].

Significant effort has been invested in developing computer-assistive or automated staging technologies, but they have struggled to achieve human-level performance [3, 14–25]. In order for a staging system to have clinical utility it should be at least as accurate and reliable as a trained human scorer. Therefore, a practical non-inferiority threshold for staging algorithms is an overall agreement of 82.0% (Cohen's kappa = 0.76), which is the overall inter-rater agreement between trained scorers at eight European centers using the 2007 AASM PSG scoring rules [5].

Machine learning is a field of computer science where classifiers discover novel patterns within a dataset without the traditional explicit encoding of all rules. Because PSG data are complex, different machine learning methods for detecting sleep stages have been trialed over the last 20 years. Published models have used hand-tuned feature extraction techniques such as spectral power, time domain analysis, and time–frequency domain (wavelet) analysis [26–29]. Other systems use fuzzy logic, support vector machines, hidden Markov models, or artificial neural networks [30–38]. Most of these systems do not achieve human-level inter-rater agreement or are tested against a small set of preselected, high-quality PSGs that do not reflect realistic testing environments. Few have been validated against large clinical datasets. In recent years, deep neural networks have rapidly found favor for signal analysis. They have proven to be remarkably robust in developing classifier systems for noisy, "real-world" datasets: the type of data represented by PSGs.

A standard neural network consists of a number of simple connected processors called neurons that mathematically transform an input signal into an output. The relative strength, or weight, of each neuron is iteratively adjusted during model training to maximize the accuracy between the network output and the expected value. Deep neural networks have many layers of neurons, where the output of one layer provides the input to the next layer, enabling discovery of nonlinear and hierarchical relationships within the data. Convolutional neural networks (CNNs) emphasize patterns in close spatial proximity and are well suited to problems in the image classification and recognition space [39, 40]. Recurrent neural networks function well with information contained in sequences such as natural language, where the next word or character depends on the immediately preceding data [41]. PSGs are well suited for convolutional and recurrent processing methodologies as they consist of spatially and temporally related signal data. For example, a k-complex may signal onset of N2 sleep, even though subsequent electromyogram (EMG) data may be low-amplitude mixed-frequency data visually identical to N1.

The increase in available computing power and publicly available PSG datasets over the last several years has brought the era of Big Data and machine learning to sleep medicine and made deep neural network processing of PSGs feasible [42, 43]. Successful development of a reliable and accurate automated scoring system using machine learning will ease the burden of PSG scoring and will reduce sleep staging inter-rater variability that affects Sleep Medicine research and clinical practice.

Methods

This study was designed as a retrospective analysis of PSG data collected through several multicenter cohort studies made available through the National Sleep Research Resource (NSRR) [43–45]. The study design was approved by the Vanderbilt University Medical Center Institutional Review Board (#171186) and data access was approved by the NSRR.

Study datasets

A deep neural network model was trained and tested on 5804 Type II PSGs from multiple centers containing patients with and without sleep-disordered breathing collected for the Sleep Heart Health Study (SHHS; Table 1) [43–45]. Two additional unrelated datasets available through the NSRR were used to test the generalizability of the model: the Study of Osteoporotic Fractures (SOF) and the Osteoporotic Fractures in Men study (MrOS; Table 2).

PSG data

All PSG files were downloaded in the European Data Format which contained the raw time series data of physiologic signals from each PSG as well as human scored sleep stages and apneic events. For the training phase, 5213 PSGs were randomly selected from the SHHS dataset, providing 42 560 hours of sleep data in 5 107 200 30-second epochs. PSGs in all three datasets were recorded as Type II unattended home studies previously scored using modified RK criteria [43–45]. PSG signal data and sleep stage labeling (Wake, N1, N2, N3, N4, or REM) were extracted from each study cohort. RK stages 3 and 4 were combined into a single stage N3 label to more closely align with modern AASM scoring conventions and to aid comparison with previously published literature. The model was trained and tuned using 90% of the SHHS visit 1 data (5213 patients). A 10% holdout set (580 patients) was taken and set aside to validate the model.

Input data and feature selection

Signal data from the electroencephalogram (EEG), EMG, and electrooculogram (EOG) PSG channels were extracted for model analysis. The Type II PSGs across all three cohorts were recorded

Table 1. Sleep Heart Health Study summary statistics

Category	Mean	Median	Min, max
Age	63.1	63	[39, 90]
Body mass index	28.2	27.5	[18, 50]
Apnea–hypopnea index	17.9	13.2	[0, 161.8]
Sleep time (minutes)	359.8	367.0	[34.5, 519]

using a single central (C3) EEG channel. Sampling rates across data channels from SOF and MrOS were down- or up-sampled as indicated to match corresponding baseline data sampling rates from SHHS.

Two different methodologies for feature representation were tested. In the first method, raw PSG signal data were provided directly as input to the network in per-epoch units and tested under various model architectures. In the second method, short-time Fourier transforms were used to generate a spectrogram for each epoch and then provided to the model as the input. Spectrograms were generated using 2-second sub-epochs formed by a Tukey window with 25% of the window inside the tapered cosine region (Figure 1). Signal normalization and filter signal preprocessing methods (median, finite impulse response, and infinite impulse response filters) were tested to evaluate the impact of noise and artifact reduction.

All data preprocessing was performed using the signal module in the python packages SciPy and scikit-learn. Model development was performed using Keras on a TensorFlow backend.

Model architecture

Convolutional and recurrent network layers were used to take advantage of the temporally linked, sequential construction of PSG data. Convolutional layers were generated to evaluate the co-occurrence of signal patterns within one-dimensional PSG data channels or co-occurrence of frequencies within single spectrograms. Recurrent layers were designed to take advantage of the temporal relationships in the data such as epochs of equivalent stage occurring in sequence. The deep neural network combined recurrent and convolutional structures to evaluate input spectrograms generated from the raw data (Figure 2). Multiple combinations of dense, convolutional, and recurrent layers were tested against the training set in the network architecture (Supplementary Appendix A).

Model tuning

Deep neural networks contain tunable hyperparameters (i.e. number of layers, number of units in each layer, number of filters in convolutional layers). A set of parameter search spaces were defined for each hyperparameter, and the best combination of hyperparameters were found using the python package hyperopt with a random search algorithm for parameter tuning [46]. Multiple hyperparameter configurations were evaluated using the training set.

Model evaluation

Model performance was evaluated with accuracy, F1-score, and Cohen's kappa. Weighted and unweighted accuracy and F1-score were calculated to assess the effect of sleep stage class imbalances in the data. Weighted accuracy was calculated as the average of the per-class stage accuracies. Because the "ground truth" comparators are human-tagged PSG events with their own level of inter-rater reliability, model agreement was also assessed using inter-rater agreement statistics (Cohen's kappa). Transition epoch F1-scores were calculated as scoring

Table 2. Summary of datasets used in study

Dataset	Polysomnography studies (*n*)	Study population	W (%)	N1 (%)	N2 (%)	N3 (%)	R (%)
SHHS	5793	Adults aged 40 years and older	28.8	3.7	40.9	12.6	13.9
MrOS	2907	Men 65 years or older	46.1	3.7	33.9	5.8	10.6
SOF	461	Women ages 65–89 years	41.9	2.9	32.5	11.9	10.7

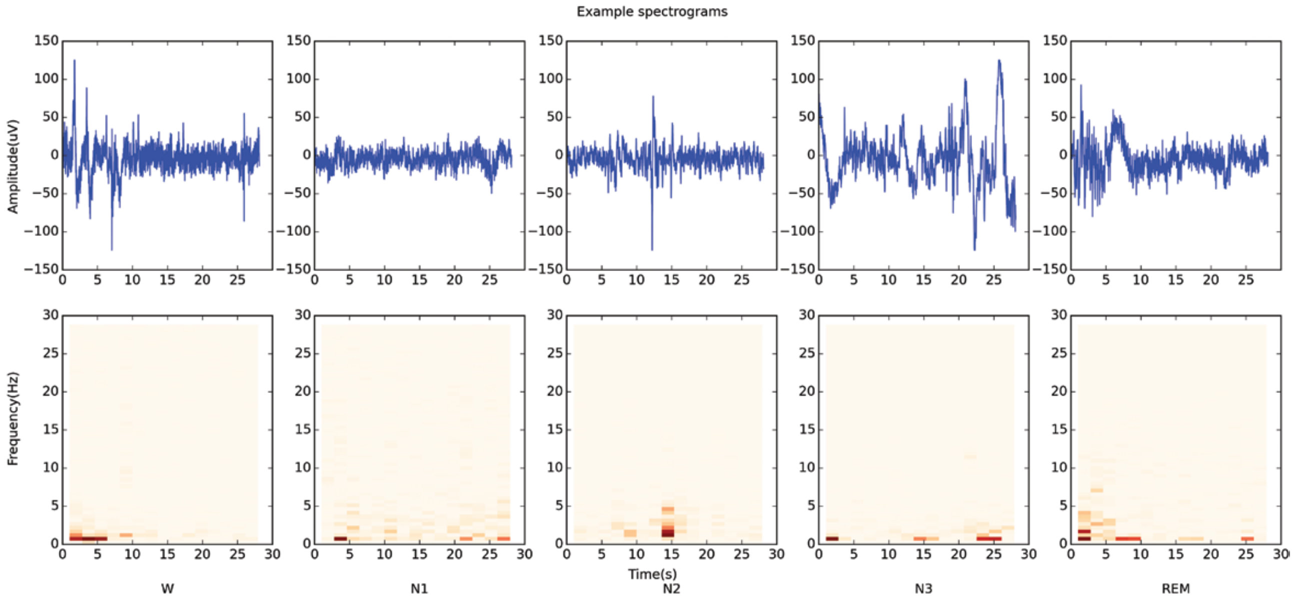

Figure 1. Representative raw data sample from each sleep stage with associated spectrogram.

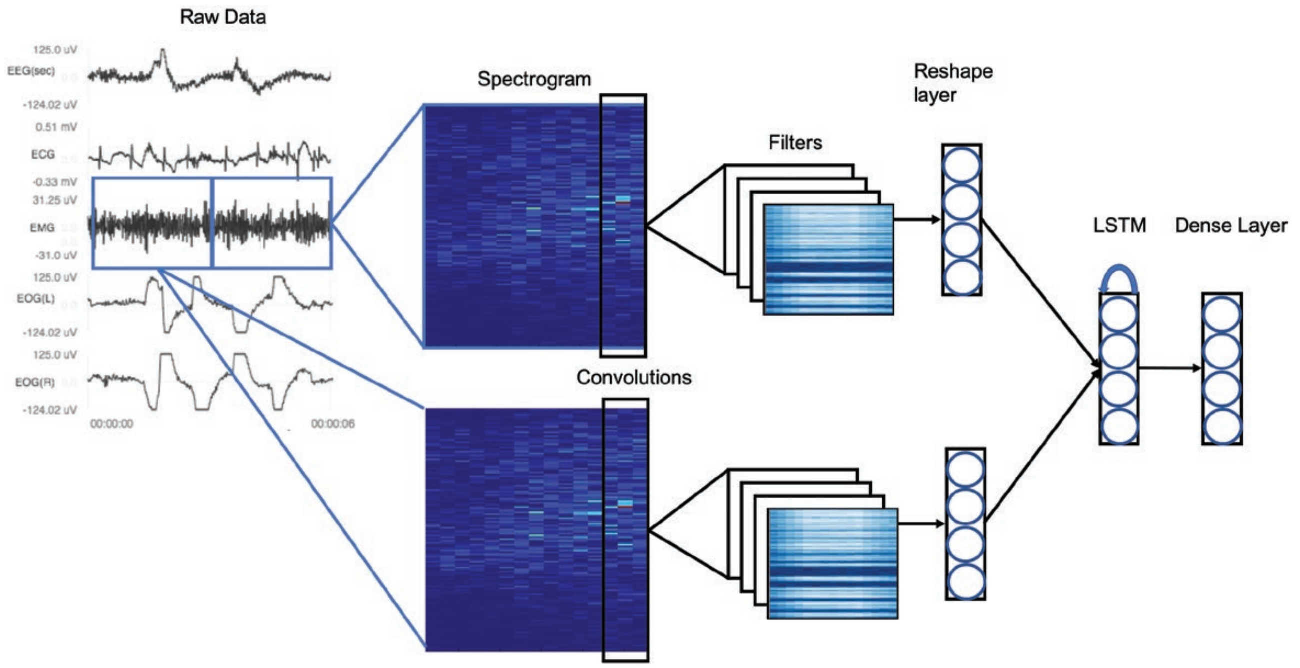

Figure 2. Simplified example model architecture for one data channel. LSTM = long short-term memory layer.

agreement is known to degrade during transition from one stage of sleep to another. Transition stages account for approximately 0.5% of the data, but were evaluated as they potentially convey physiologically relevant information.

Transfer learning

Generalizability was assessed using the SOF and MrOS datasets. These studies were conducted in different environments with various types of acquisition hardware and on different patient populations than SHHS.

Model performance was additionally evaluated on subsets of the SHHS population with mild, moderate, and severe obstructive sleep apnea (OSA) to demonstrate model transferability between patients with different degrees of sleep-disordered breathing. A separate model was also trained and tested on only severe patients to demonstrate validity even when restricted to a subset of studied patients.

Results

The optimal sleep staging model's architecture consisted of a combination of separate networks for each signal channel. Spectrograms of each channel were fed into convolutional layers that examined the proximal relationships of the frequencies in time as well as recurrent layers that examined the sequential relationships of epochs (Table 3). The subnetworks for each signal channel were combined into two dense layers feeding into a final softmax output layer used to generate discrete stage predictions for each epoch.

Model testing

The SHHS dataset was split into a 90% training and 10% holdout set. The training set was further split into training and validation sets, which were used to train the model, select the optimal deep learning architecture (Supplementary Appendix A), and tune the model hyperparameters (Supplementary Appendix B). Model training required approximately 48 hours on an Nvidia GTX Titan X GPU. A learning curve plateauing around 1 000 000 training epochs demonstrated that the dataset was sufficiently large (Figure 3). Testing on the holdout set required approximately 30 minutes.

Table 3. Base model architecture per data channel

Layer	Layer type	Size	Output size
Input			(2, 1, 129, 16)
C1	Convolutional	(32, 64, 3)	(2, 32, 66, 14)
C2	Convolutional	(32, 64, 3)	(2, 32, 2, 12)
P1	Max pooling	(2, 2)	(2, 32, 1, 6)
R1	Reshape		(2, 192)
L1	Long short-term memory	256	256
D1	Dense	512	512

Model evaluation

Signal preprocessing methods were tested on the raw input signal. No significant improvement in accuracy or F1-scores were found using normalization or filters, so signal preprocessing was not used in the final pipeline (data not shown). Multiple model architectures were tested on the SHHS dataset. The first model was a simple baseline Markov chain that predicted the next stage based on overall stage transition probabilities measured directly from SHHS. Because stages commonly occur in long chains with relatively rare transitions, this model has a high F1-score, but low transition F1-score. Following this baseline model, a CNN was tested against raw PSG data, followed by separate CNN and long short-term memory (LSTM) models on the spectrogram data, and finally a combination of CNN + LSTM, which yielded the best performance (Figure 4).

The optimal neural network model was composed of spectrograms in the input layer feeding into CNN layers and an LSTM layer to achieve a weighted F1-score of 0.87 and Cohen's Unweighted kappa of K = 0.82, higher than that of human agreement found in literature (K = 0.76).

A confusion matrix was generated for model performance against all tested epochs (Figure 5) as well as transition epochs (Figure 6). When considering all epochs, the model scored Wake, N1, N2, N3, and REM stages correctly 92%, 37%, 91%, 77%, and 88% of the time, respectively. During transition epochs correct staging was scored for Wake, N1, N2, N3, and REM 75%, 44%, 79%, 54%, and 88% of the time, respectively. Table 4 compares staging accuracy of this model to others published in the literature using the class imbalances present in the underlying dataset. Table 5 permits comparison to other models in the literature that used methods to balance the classes such that all classes contribute equally in model training. Figure 7 demonstrates agreement between a trained scorer and the automated scoring model in one example PSG hypnogram.

Performance on cohorts with and without sleep-disordered breathing

The model performs similarly on subsets of the holdout set with different apnea severity (Table 6). A model trained and tested on severe OSA patients only achieved an unweighted F1-score of 0.846, similar to the model trained on heterogeneous data.

Transfer learning

After training on SHHS data, model generalizability was tested against two additional NSRR datasets. The microvolt mean and SD of each included data channel was significantly different between studies, suggesting different signal architectures between datasets (Table 7).

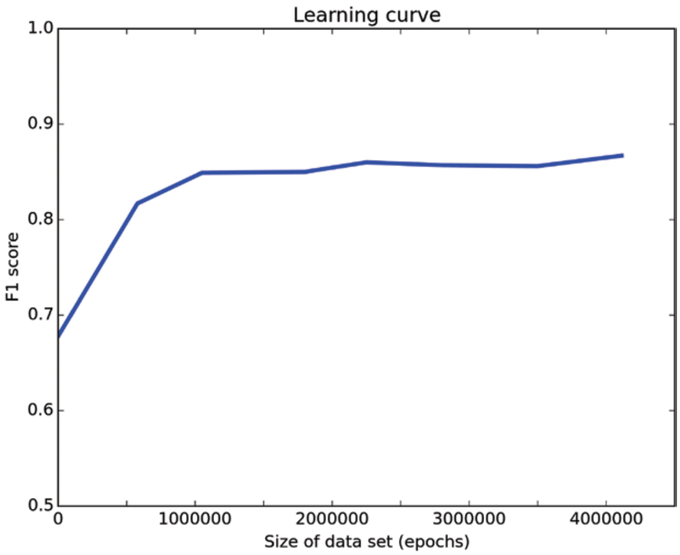

Figure 3. The deep neural network model learning curve begins to plateau after training on approximately 1 000 000 epochs.

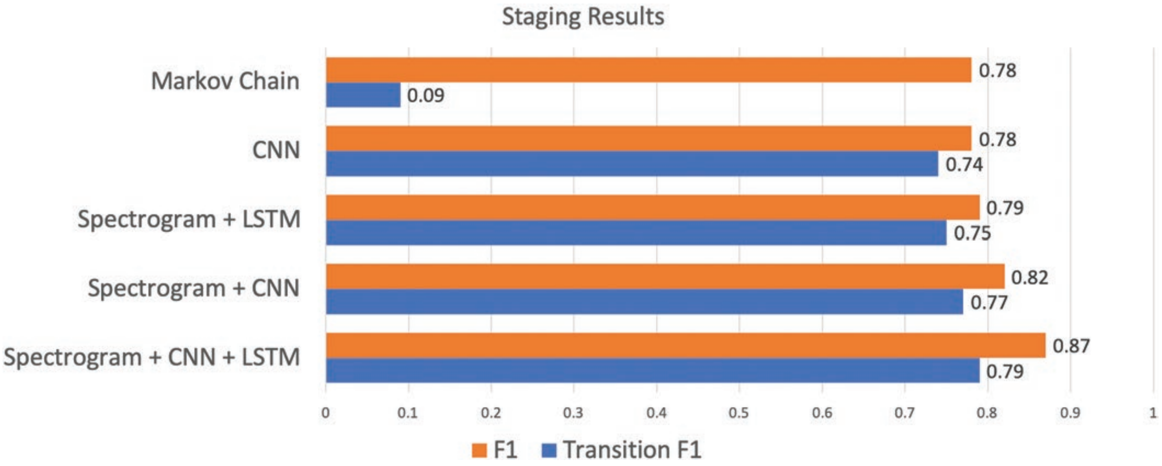

Figure 4. Model performance under various architectures against the SHHS dataset. CNN = convolutional neural network; LSTM = long short-term memory

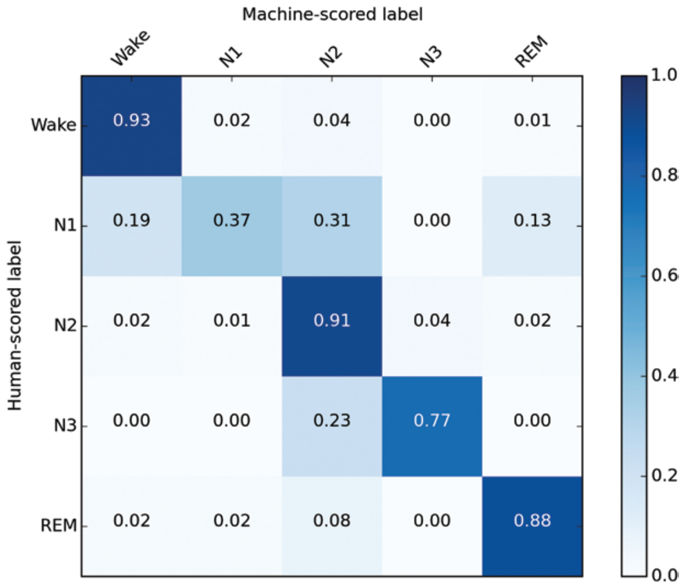

Figure 5. Confusion matrix for all epochs.

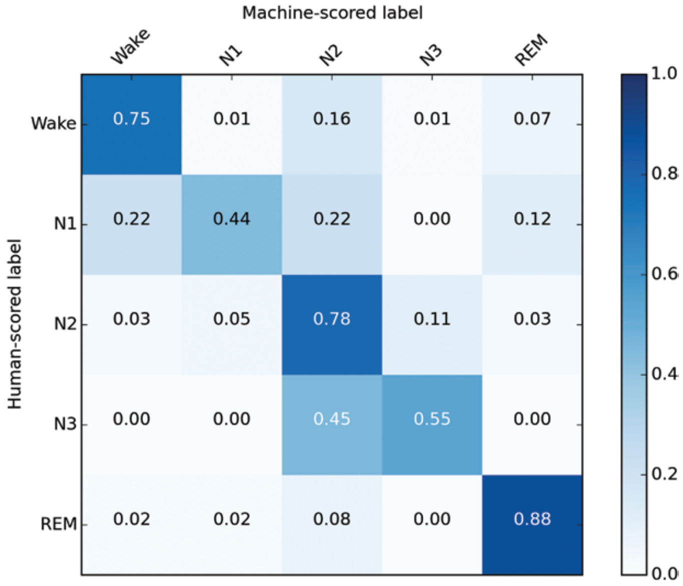

Figure 6. Transition epoch confusion matrix.

F1-score and Cohen's kappa scores on the MrOS and SOF datasets demonstrated moderate-to-strong inter-rater agreement between the model and trained scorers depending on the selected testing data and achieved high performance in the balance of precision and recall on sleep staging (Table 8).

Discussion

The deep learning model presented here automatically predicts sleep stage with moderate-to-strong agreement compared with expert human scorers across multiple datasets. The optimal model used input consisting of spectrograms derived from the EEG, EMG, and EOG channels passed to a deep learning architecture with convolutional and recurrent layers. A learning curve demonstrated that sufficient data was available to train the model well. The model performs comparably or better than other models reported in literature and, when tested against studies with structure similar to the underlying training dataset, meets or exceeds the accepted benchmark of K = 0.76 between trained human scorers.

Spectrograms are used to represent the data provided to the model in the form of dimensionally reduced input that retains important information for sleep stage classification. The Fourier transforms used to generate spectrograms organized PSG data into component frequencies more easily compared across different platforms than raw signal data, which contains baseline signal noise and variation due to different recording environments and hardware. Spectrogram construction also aided

Table 4. Performance of class imbalanced model compared to other studies

Study	Sample size (studies)	Evaluation split	W Accuracy	N1 Accuracy	N2 Accuracy	N3 Accuracy	REM Accuracy	Overall Accuracy	Balanced Accuracy	Cohen's kappa
Biswal *et al.* [47]	10 000	Train–validation–test	84.5%	56.2%	88.4%	85.4%	92%	85.8%	81.3%	0.795
Sors *et al.* [48]	5793	Training–validation–test	91%	35%	89%	85%	86%	87%	77.2%	0.81
Sharma *et al.* [49]	100	10-fold-CV	95%	17%	76%	57%	36%	91.7%	56.5%	N/A
Proposed model	5793	Train–validation–test	92%	37%	91%	77%	88%	87%	77%	0.82

Table 5. Performance of class balanced model compared to other studies

Study	Sample size (studies)	Evaluation split	W Accuracy	N1 Accuracy	N2 Accuracy	N3 Accuracy	REM Accuracy	Overall Accuracy	Balanced Accuracy	F1-Score	Cohen's kappa
Supratak *et al.* [50]	62	31-fold cross validation	87.3%	43.5%	90.5%	77.1%	80.9%	86.2%	75.9%	0.817	0.8
Tsinalis *et al.* [51]	40	20-fold cross validation	70%	60%	73%	91%	74%	82%	74%	0.81	N/A
Chambon *et al.* [52]	62	5-fold cross validation	85%	52%	77%	91%	83%	79%	77.6%	0.72	N/A
Proposed model	5793	Train–validation–test	91%	46%	89%	77%	88%	86%	78%	0.81	0.82

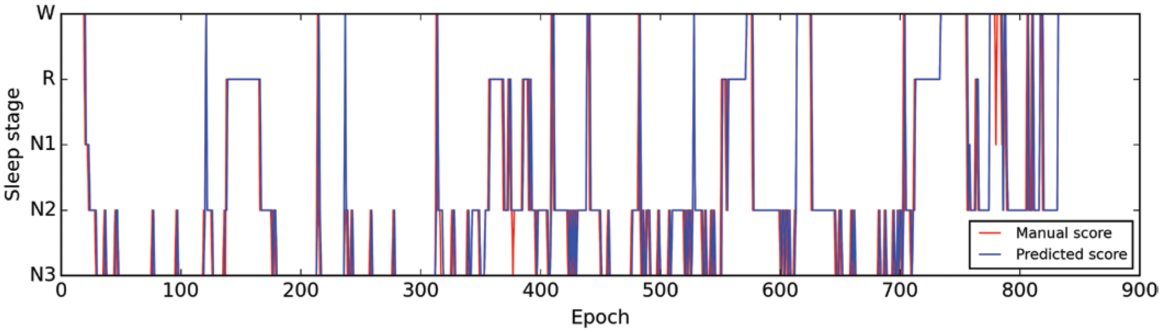

Figure 7. Example output hypnogram of a PSG scored by the model overlaid on the human manual scoring.

network throughput as the volume of input data were reduced without significant loss of key signal information.

Preprocessing raw signal data for noise and artifact reduction did not significantly affect classification results in preliminary testing. Prior performance analyses have demonstrated that deep learning models become more robust when trained on noisy data [53], and we suspect that training on raw, unprocessed data may be advantageous for accuracy and transferability when testing across clinical datasets as well. Noisy input data are hypothesized to improve the robustness of deep learning models by stabilizing against distortions in the input [54]. Networks trained on unprocessed data are better able to handle noise arising in unseen data. By training neural networks on unprocessed data, the need for preprocessing in new data is reduced and a greater proportion of relevant signal data can be preserved for analysis.

PSGs have significant class imbalances between stage types due to the natural asymmetric distribution of sleep stages. The SHHS dataset is no exception, with large differences in representation between several of the stages. Accounting for class imbalances by overrepresenting minority classes (such as N1) can improve single class accuracy, but often at the expense of larger classes. For instance, in SHHS N1 is only 3.7% of the dataset, whereas N2 is 40.9%. The model presented here scored 31% of N1 and 91% of N2 epochs correctly with an overall accuracy of 87% when the native class imbalances are not adjusted.

When N1 was oversampled to balance class representation, accuracy of N1 increased to 45% at the expense of other stages, such as N2, which decreased to an accuracy of 88%. Class balancing decreased overall model scoring accuracy to 86%. Class imbalances also complicate comparison of performance metrics between published models. We believe that preserving native class imbalances best represents how the model would perform in a production setting. However, performance metrics for models trained on natural as well as balanced class distributions are provided in order to facilitate comparison with previously published models (Tables 4 and 5).

Accuracy in N1 scoring is worse than other sleep stages for this model, consistent with other published models [48–52]. This may be an artifact of PSG scoring rules, which allow for low-amplitude mixed pattern EEG signals identical to N1 to be scored as N2 if the preceding stage was also scored as N2. These rules, along with the large class imbalances between N1 and N2, likely compromise N1 accuracy.

Other issues may complicate scoring accuracy, such as patient movement artifacts contaminating W and N1 stages. Unlike many

Table 6. SHHS model performance on patient subgroups of varying obstructive sleep apnea severity

Testing cohort	F1	Epochs (*N*)
All	0.872	621 794
Normal (AHI < 5)	0.871	132 742
Mild (5 < AHI < 15)	0.864	262 426
Moderate (15 < AHI < 30)	0.853	168 074
Severe (AHI > 30)	0.841	58 552

AHI = apnea–hypopnea index.

Table 7. Mean and SD of the channels for each dataset

Channel	SHHS	MrOS	SOF
EEG (uV)	−0.39 ± 30.31	2.5 ± 38.08*	−8.87 ± 43.02*
EMG (uV)	0.54 ± 9.68	−1.06 ± 58.49*	10.05 ± 34.47*
EOG(L) (uV)	−3.57 ± 30.60	−12.5 ± 49.28*	−9.81 ± 35.60*
EOG(R) (uV)	−4.19 ± 31.36	3.33 ± 50.81*	5.32 ± 41.37*

*indicates significant difference from SHHS data at $p < 0.05$.

Table 8. Generalizability of the SHHS model to novel datasets

Model	F1-score (weighted)	Cohen's kappa
Training data: SHHS Testing data: SHHS	0.87	0.82
Training data: SHHS Testing data: MrOS	0.79	0.70
Training data: SHHS Testing data: SOF	0.77	0.68
Training data: MrOS Testing data: SHHS	0.69	0.56
Training data: SOF Testing data: SHHS	0.66	0.53

other published works, this model was not trained on a curated set of high-quality PSGs and contains studies partially contaminated by signal and motion artifacts. Contaminated epochs scored by humans theoretically contain enough signal information that they should be of value in training a machine learning algorithm that will be exposed to similar data in a production environment. The inclusion of this more ambiguous data may create systemic difficulties in scoring W and N1 in the same way that it would degrade inter-rater agreement between human scorers. To this point, Younes *et al.* [13] recently found an intra-class correlation coefficient of 0.69 (range: 0.30–0.86) in N1 scoring, suggesting only poor to moderate agreement between trained human scorers.

This model presented in this work has several strengths. It meets or exceeds performance of other published works. A large and diverse training dataset increased transferability, demonstrated across several other large datasets. Significant differences existed in mean microvolt channel levels across the tested datasets (Table 6), suggesting significant underlying differences in dataset structure due to differences in recording hardware, environment, study populations, or other variables. Despite these differences, the model presented here could be trained on one dataset and still perform with moderate-to-strong agreement on other datasets (MrOS F1 = 0.78, K = 0.68 and SOF F1 = 0.68, K = 0.55). The model also performed similarly on cohorts composed of subjects with varying degrees of sleep disordered breathing, with F1-scores ranging from 0.841 to 0.872, suggesting that sleep-disordered breathing does not significantly affect sleep stage classification patterns for the model. In comparison, a model trained only on patients with severe sleep apnea and tested on the same cohort performs only slightly better than one trained on all patients, demonstrating model transferability between different disease populations. Taken together, the transferability properties illustrated here suggest that automated deep learning classifiers have the potential for use in different clinical sleep laboratory environments without complete retraining on local data.

Few other studies test models on PSGs collected from a variety of recording environments and hardware platforms. Patanaik *et al.* [55] did so, demonstrating generalizability by testing against two novel datasets with inter-rater agreement of K = 0.740 and K = 0.597. However, their reported outcomes (accuracy) were obtained from model training data instead of separate holdout data, limiting inner-dataset comparability to the work presented here. The kappa values are also not directly comparable to our inter-rater agreement of K = 0.70 and K = 0.56. The datasets in Patanaik *et al.* were acquired using the same framework and pipeline, whereas the external test datasets presented here were acquired on a variety of different hardware platforms that were then down- or up-sampled to match SHHS dataset frequencies. Both studies demonstrate comparable performance on external datasets that the models were not trained on, demonstrating transferability.

This work is not without limitations. The datasets examined here are composed of Type II PSGs recorded in subject home environments with a limited, single EEG channel montage. Generalizability to more common Type I or Type III PSGs could not be evaluated; however, we suspect that training the model with additional EEG signals available in Type I PSGs would likely yield performance improvements from additional channel data. Retraining the model with additional channels while maintaining input from previously evaluated channels would be expected to improve performance, as deep neural networks generally perform better as more data is available [56]. Comparison with more limited montage datasets, such as consumer wearables using actigraphy and heart rate monitoring, is limited by the lack of large, publicly available datasets. In addition, accuracy outcomes may differ between AASM sleep staging criteria and RK staging criteria.

In conclusion, this work suggests that automated PSG scoring systems can rapidly annotate PSG files with inter-rater agreement rivaling that of trained human scorers. Future work will require institutions and interested stakeholders to make available large libraries of high-quality datasets using modern scoring criteria in order for data scientists to develop robust, generalizable scoring models.

Supplementary material

Supplementary material is available at *SLEEP* online.

Funding

This work was supported by the NLM training grant 4T15LM 7450-15 from the National Institutes of Health.

Conflict of interest statement. None declared.

References

1. Rechtschaffen A, *et al*. *A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects*. Los Angeles, CA: National Institutes of Health Publication no. 204; 1968.
2. Iber C. *et al*. *The AASM manual for the scoring of sleep and associated events: Rules, terminology and technical specifications*. Vol. **1** Westchester, IL: American Academy of Sleep Medicine; 2007.
3. Malhotra A, *et al*. Performance of an automated polysomnography scoring system versus computer-assisted manual scoring. *Sleep*. 2013;**36**(4):573–582.
4. Silber MH, *et al*. The visual scoring of sleep in adults. *J Clin Sleep Med*. 2007;**3**(2):121–131.
5. 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.
6. Collop NA. Scoring variability between polysomnography technologists in different sleep laboratories. *Sleep Med*. 2002;**3**(1):43–47.
7. Loredo JS, *et al*. Night-to-night arousal variability and interscorer reliability of arousal measurements. *Sleep*. 1999;**22**(7):916–920.
8. Norman RG, *et al*. Interobserver agreement among sleep scorers from different centers in a large dataset. *Sleep*. 2000;**23**(7):901–908.
9. Bliwise D, *et al*. Measurement error in visually scored electrophysiological data: respiration during sleep. *J Neurosci Methods*. 1984;**12**(1):49–56.
10. Lord S, *et al*. Interrater reliability of computer-assisted scoring of breathing during sleep. *Sleep*. 1989;**12**(6):550–558.
11. Whitney CW, *et al*. Reliability of scoring respiratory disturbance indices and sleep staging. *Sleep*. 1998;**21**(7):749–757.
12. Drinnan MJ, *et al*. Interobserver variability in recognizing arousal in respiratory sleep disorders. *Am J Respir Crit Care Med*. 1998;**158**(2):358–362.
13. Younes M, *et al*. Reliability of the American Academy of Sleep Medicine rules for assessing sleep depth in clinical practice. *J Clin Sleep Med*. 2018;**14**(2):205–213.
14. Schaltenbrand N, *et al*. Sleep stage scoring using the neural network model: comparison between visual and automatic analysis in normal subjects and patients. *Sleep*. 1996;**19**(1):26–35.
15. Anderer P, *et al*. An E-health solution for automatic sleep classification according to Rechtschaffen and Kales: validation study of the Somnolyzer 24 × 7 utilizing the Siesta database. *Neuropsychobiology*. 2005;**51**(3):115–133.
16. Berthomier C, *et al*. Automatic analysis of single-channel sleep EEG: validation in healthy individuals. *Sleep*. 2007;**30**(11):1587–1595.
17. Anderer P, *et al*. Computer-assisted sleep classification according to the standard of the American Academy of Sleep Medicine: validation study of the AASM version of the Somnolyzer 24 × 7. *Neuropsychobiology*. 2010;**62**(4):250–264.
18. Fraiwan L, *et al*. Classification of sleep stages using multi-wavelet time frequency entropy and LDA. *Methods Inf Med*. 2010;**49**(3):230–237.
19. Liang SF, *et al*. A rule-based automatic sleep staging method. *J Neurosci Methods*. 2012;**205**(1):169–176.
20. Lajnef T, *et al*. Learning machines and sleeping brains: automatic sleep stage classification using decision-tree multi-class support vector machines. *J Neurosci Methods*. 2015;**250**:94–105.
21. Wang Y, *et al*. Evaluation of an automated single-channel sleep staging algorithm. *Nat Sci Sleep*. 2015;**7**:101–111.
22. Punjabi NM, *et al*. Computer-assisted automated scoring of polysomnograms using the somnolyzer system. *Sleep*. 2015;**38**(10):1555–1566.
23. Hassan AR, *et al*. A decision support system for automatic sleep staging from EEG signals using tunable Q-factor wavelet transform and spectral features. *J Neurosci Methods*. 2016;**271**:107–118.
24. Younes M, *et al*. Accuracy of automatic polysomnography scoring using frontal electrodes. *J Clin Sleep Med*. 2016;**12**(5):735–746.
25. Younes M, *et al*. Performance of a new portable wireless sleep monitor. *J Clin Sleep Med*. 2017;**13**(2):245–258.
26. Agarwal R, *et al*. Computer-assisted sleep staging. *IEEE Trans Biomed Eng*. 2001;**48**(12):1412–1423.
27. Zoubek L, *et al*. Feature selection for sleep/wake stages classification using data driven methods. *Biomed Signal Process Control*. 2007;**2**(3):171–179.
28. Fraiwan L, *et al*. Automatic sleep stage scoring with wavelet packets based on single EEG recording. *World Acad Sci Eng Technol*. 2009;**54**(3):485–488.
29. Bajaj V, *et al*. Automatic classification of sleep stages based on the time-frequency image of EEG signals. *Comput Methods Programs Biomed*. 2013;**112**(3):320–328.
30. Chapotot F, *et al*. Automated sleep–wake staging combining robust feature extraction, artificial neural network classification, and flexible decision rules. *Int J Adapt Control Signal Process*. 2010;**24**(5):409–423.
31. Figueroa Helland VC, *et al*. Investigation of an automatic sleep stage classification by means of multiscorer hypnogram. *Methods Inf Med*. 2010;**49**(5):467–472.
32. Jo HG, *et al*. Genetic fuzzy classifier for sleep stage identification. *Comput Biol Med*. 2010;**40**(7):629–634.
33. Güneş S, *et al*. Efficient sleep stage recognition system based on EEG signal using k-means clustering based feature weighting. *Expert Syst Appl*. 2010;**37**(12):7922–7928.
34. Doroshenkov LG, *et al*. Classification of human sleep stages based on EEG processing using hidden Markov models. *Biomed Eng*. 2007;**41**(1):25–28.
35. Dong J, *et al*. Automated sleep staging technique based on the empirical mode decomposition algorithm: a preliminary study. *Advances in Adaptive Data Analysis*. 2010;**2**(02):267–276.
36. Koley B, *et al*. An ensemble system for automatic sleep stage classification using single channel EEG signal. *Comput Biol Med*. 2012;**42**(12):1186–1195.
37. Hsu Y-L, *et al*. Automatic sleep stage recurrent neural classifier using energy features of EEG signals. *Neurocomputing* 2013;**104**:105–114.
38. Krakovská A, *et al*. Automatic sleep scoring: a search for an optimal combination of measures. *Artif Intell Med*. 2011;**53**(1):25–33.
39. Krizhevsky A, *et al*. ImageNet classification with deep convolutional neural networks. *NIPS*. In: Advances in neural information processing systems. Dec 3-8, 2012; Curran Associates, Red Hook, NY. 1106–1114.
40. Cecotti H, *et al*. Convolutional neural networks for P300 detection with application to brain-computer interfaces. *IEEE Trans Pattern Anal Mach Intell*. 2011;**33**(3):433–445.
41. Graves A, *et al*. Speech recognition with deep recurrent neural networks. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). May 26-31, 2013; IEEE, Piscataway, NJ. 2013: 6645–6649.

42. Budhiraja R, *et al*. The role of big data in the management of sleep-disordered breathing. *Sleep Med Clin*. 2016;**11**(2):241–255.
43. Dean DA II, *et al*. Scaling up scientific discovery in sleep medicine: the National Sleep Research Resource. *Sleep*. 2016;**39**(5):1151–1164.
44. Quan SF, *et al*. The Sleep Heart Health Study: design, rationale, and methods. *Sleep*. 1997;**12**:1077–1085.
45. Redline S, *et al*. Methods for obtaining and analyzing unattended polysomnography data for a multicenter study. Sleep Heart Health Research Group. *Sleep*. 1998;**21**(7):759–767.
46. Pan SJ, *et al*. A survey on transfer learning. *IEEE Trans Knowl Data Eng*. 2010;**22**(10):1345–1359.
47. Biswal S, *et al*. SLEEPNET: automated sleep staging system Via deep learning. arXiv. 2017: 1–17. https://arxiv.org/pdf/1707.08262.pdf. Accessed January 1, 2018.
48. Sors A, *et al*. A convolutional neural network for sleep stage scoring from raw single-channel EEG. *Biomed Signal Process Control*. 2018;**42**:107–114.
49. Sharma M, *et al*. An accurate sleep stages classification system using a new class of optimally time-frequency localized three-band wavelet filter bank. *Comput Biol Med*. 2018;**98**:58–75.
50. 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.
51. Tsinalis O, *et al*. Automatic sleep stage scoring with single-channel EEG using convolutional neural networks. arXiv. 2016; 1–10. https://arxiv. org/abs/1610.01683. Accessed January 1, 2018.
52. Chambon S, *et al*. A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series. *IEEE Trans Neural Syst Rehabil Eng*. 2018;**26**(4):758–769.
53. Nazaré TS, *et al*. Deep convolutional neural networks and noisy images. InIberoamerican Congress on Pattern Recognition; Nov 7, 2017 (pp. 416–424). Springer, Cham.
54. Stephan Z, *et al*. Improving the robustness of deep neural networks via stability training. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Jun 26, 2016 – Jul 1, 2016; IEEE, Piscataway, NJ. 4480–4488.
55. Patanaik A, *et al*. An end-to-end framework for real-time automatic sleep stage classification. *Sleep*. 2018;**41**(5):1–11.
56. Sun C, *et al*. *Revisiting Unreasonable Effectiveness of Data in Deep Learning Era*. In: ICCV. Oct 22-29, 2017; IEEE, Piscataway, NJ.