SLEEPJ, 2020, 1–12

doi: 10.1093/sleep/zsaa112
Advance Access Publication Date: 1 June 2020
Original Article

ORIGINAL ARTICLE

Expert-level automated sleep staging of long-term scalp electroencephalography recordings using deep learning

Maurice Abou Jaoude[1,], Haoqi Sun[1,], Kyle R. Pellerin[1], Milena Pavlova[2], Rani A. Sarkis[2], Sydney S. Cash[1], M. Brandon Westover[1,†], and Alice D. Lam[1,*,†]

[1]Department of Neurology, Massachusetts General Hospital and Harvard Medical School, Boston, MA and
[2]Department of Neurology, Brigham and Women's Hospital and Harvard Medical School, Boston, MA

[†]These authors are co-senior authors to this work.

[*]Corresponding author. Alice D. Lam, Department of Neurology, Massachusetts General Hospital 55 Fruit Street, WACC 735, Boston, MA 02114. Email: lam.alice@mgh.harvard.edu.

Abstract

Study Objectives: Develop a high-performing, automated sleep scoring algorithm that can be applied to long-term scalp electroencephalography (EEG) recordings.

Methods: Using a clinical dataset of polysomnograms from 6,431 patients (MGH–PSG dataset), we trained a deep neural network to classify sleep stages based on scalp EEG data. The algorithm consists of a convolutional neural network for feature extraction, followed by a recurrent neural network that extracts temporal dependencies of sleep stages. The algorithm's inputs are four scalp EEG bipolar channels (F3-C3, C3-O1, F4-C4, and C4-O2), which can be derived from any standard PSG or scalp EEG recording. We initially trained the algorithm on the MGH–PSG dataset and used transfer learning to fine-tune it on a dataset of long-term (24–72 h) scalp EEG recordings from 112 patients (scalpEEG dataset).

Results: The algorithm achieved a Cohen's kappa of 0.74 on the MGH–PSG holdout testing set and cross-validated Cohen's kappa of 0.78 after optimization on the scalpEEG dataset. The algorithm also performed well on two publicly available PSG datasets, demonstrating high generalizability. Performance on all datasets was comparable to the inter-rater agreement of human sleep staging experts (Cohen's kappa ~ 0.75 ± 0.11). The algorithm's performance on long-term scalp EEGs was robust over a wide age range and across common EEG background abnormalities.

Conclusion: We developed a deep learning algorithm that achieves human expert level sleep staging performance on long-term scalp EEG recordings. This algorithm, which we have made publicly available, greatly facilitates the use of large long-term EEG clinical datasets for sleep-related research.

Statement of Significance

Long-term scalp electroencephalography (EEG) recordings are increasingly used for clinical purposes and capture a significant amount of sleep. Repurposing large clinical EEG datasets for sleep research would be advantageous, but a major barrier is that sleep staging is not typically available for clinical EEG recordings. High-performing, automated sleep staging algorithms have been developed for polysomnograms, but these algorithms require polysomnogram-specific inputs and thus cannot be applied to most long-term scalp EEG datasets. Here, we developed a sleep staging algorithm that can be applied to long-term scalp EEG data from many different clinical settings and that performs at the level of human experts. This publicly available algorithm will facilitate the use of large clinical EEG datasets for "big data" sleep research.

Key words: sleep staging; deep learning; EEG; machine learning; big data

Submitted: 4 December, 2019; **Revised:** 20 March, 2020

Introduction

Long-term scalp EEG recordings (typically > 12 h in duration) are widely used for clinical evaluation of patients with confirmed or suspected epilepsy [1] as these studies are more sensitive than "routine EEGs" (typically 30–60 min in duration) in capturing paroxysmal electrical abnormalities such as epileptiform discharges and seizures [2, 3], as well as capturing behavioral or other unusual spells for diagnostic purposes. These long-term scalp EEG studies, which include continuous inpatient EEG recordings performed in the epilepsy monitoring unit (EMU) and other hospital settings, and ambulatory EEG recordings performed at home, generate large amounts of EEG data, often with substantial recording time in the asleep state. This rich clinical data could potentially be leveraged for "big data" research studies on sleep, but a major barrier to using these clinical studies for sleep research is that long-term scalp EEG studies do not typically undergo manual sleep staging by an expert sleep technician. As manual sleep staging is time-consuming and labor-intensive, the lack of this information often precludes the use of large clinical datasets of long-term scalp EEG recordings for sleep research.

During the past decade, there have been many advances in the automated sleep staging of polysomnography (PSG) data [4–22]. Among the most promising advances recently has been the application of deep learning for automated sleep staging of PSGs. One of the many advantages offered by deep learning methods is their ability to automatically extract features from the data that are relevant to the classification task, without relying on developing handcrafted features. In addition, compared to traditional machine learning techniques, the performance of deep neural networks continues to improve as datasets become much larger [23]. Several groups have now shown that algorithms developed using "big data" approaches, for example, deep learning methods applied to large clinical and research PSG datasets, can achieve automated sleep staging on PSGs with performance that approaches or even exceeds human performance [24–26]. While it would be ideal to apply these PSG-based algorithms to automate sleep staging of long-term scalp EEG studies, several factors limit the translation of previously developed algorithms for this purpose. First, many of these sleep-staging algorithms use additional physiologic input channels (e.g. electrooculogram [EOG] and electromyography [EMG]) that are recorded as part of conventional PSG studies but are not typically recorded for long-term scalp EEG studies. Second, most of these algorithms use scalp EEG electrodes referenced to contralateral mastoid channels, which is the convention for PSG studies, but rarely used in long-term scalp EEG studies.

The goal of this study was to develop a high-performing, reference channel-free, automated sleep staging algorithm that could easily be applied to long-term scalp EEG recordings from many different clinical settings. Using a large clinical dataset of PSG recordings from over 6,000 patients, we trained a hybrid convolutional neural network (CNN) and recurrent neural network (RNN) to learn effective and generalizable features for sleep stage scoring. We then used transfer learning to fine-tune this algorithm specifically for use on long-term ambulatory scalp EEG recordings.

Methods

Datasets

We used two datasets for training and optimization of the deep neural network, described in further detail below (MGH–PSG dataset and Ambulatory scalpEEG dataset). Demographic and sleep characteristics for both datasets are found in Table 1. Sleep staging for both datasets was performed by expert sleep technicians in nonoverlapping 30-s epochs according to standards by the American Academy of Sleep Medicine (AASM), as one of five stages: wake (W), non-rapid eye movement (NREM) stage 1 (N1), non-REM stage 2 (N2), non-REM stage 3 (N3), and rapid eye movement (REM).

All prospective and retrospective study procedures were performed under protocols approved by our centers' Institutional Review Boards. Written informed consent was obtained from all prospectively recruited research participants. The need for informed consent was waived for retrospective analysis of the MGH–PSG dataset.

MGH–PSG dataset

This dataset consists of diagnostic, split night, and continuous positive airway pressure titrations from 6,341 patients, carried out at the Massachusetts General Hospital (MGH) sleep laboratory between the years of 2009 and 2016. The PSG EEG data includes six channels: F3-M2, F4-M1, C3-M2, C4-M1, O1-M2, and O2-M1. In total, seven sleep technicians scored records across this dataset, with each record scored by a single sleep technician.

Table 1. Demographics and general characteristics of the PSG and ambulatory EEG data

Variable	MGH–PSG dataset	ScalpEEG dataset	
		MGH–BWH EEG dataset	Biobooosti dataset
Number of recordings	7,636	94	29
Patients	6,341	93	19
Age	52.77 ± 15.43	74.24 ± 8.59	46.58 ± 16.32
% Male	59.5%	39.36%	31.58%
Recording duration (h)	8.40 ± 3.36	22.16 ± 2.19	87.15 ± 36.00
Total Sleep Time (h)	7.01 ± 3.06	6.29 ± 1.55	25.51 ± 11.14
N1 (% of sleep)	17.72 ± 13.95	14.30 ± 7.65	9.76 ± 4.72
N2 (% of sleep)	51.73 ± 12.68	59.07 ± 9.92	57.32 ± 7.53
N3 (% of sleep)	15.00 ± 10.13	1.06 ± 3.75	6.42 ± 6.75
REM (% of sleep)	15.55 ± 8.22	25.57 ± 8.77	26.51 ± 7.31

Values are average ± standard deviation.

Ambulatory scalpEEG dataset

This dataset consists of ambulatory scalp EEG recordings with standard 10–20 electrodes from 112 patients, obtained from two sources (Bioboosti and MGH–BWH EEG), described in further detail below. Sleep staging of all recordings in the ambulatory scalpEEG dataset was performed by a single sleep technician, using F3, F4, C3, C4, O1, and O2 channels in a referential montage. For the Bioboosti dataset, sleep staging also utilized dedicated EOG and EMG channels.

The first source was data from the Bioboosti study [27], an insomnia study comprised of 19 healthy participants, who had difficulty sleeping but no epilepsy or other central nervous system disease. In addition to scalp EEG electrodes, these participants also had chin EMG and EOG electrodes placed according to AASM criteria.

The second source was data from an ongoing research study on aging, epilepsy, and dementia, at MGH and Brigham and Women's Hospital (MGH–BWH EEG dataset). This dataset was comprised of 93 participants, including 40 cognitively normal healthy elderly participants, 31 participants with amnestic mild cognitive impairment due to probable Alzheimer's disease, 16 with mild dementia due to probable Alzheimer's disease, and 6 with late-onset epilepsy of unclear etiology. Among the participants with probable Alzheimer's disease, nine also had a history of epilepsy related to Alzheimer's disease. Two board-certified epileptologists (A.D.L. and R.A.S.) independently reviewed all scalp EEGs and came to a consensus on the background abnormalities present in each recording, including focal slowing, diffuse slowing, and sporadic epileptiform discharges.

Validation datasets

We also tested our algorithm on two publicly available PSG datasets with sleep scoring [28, 29]: (1) the Home Positive Airway Pressure (homePAP) study [30] and (2) the Apnea, Bariatric surgery, and CPAP (ABC) study [31]. The homePAP dataset consisted of recordings from 243 patients undergoing lab-based PSG at 7 different sleep centers. The ABC dataset consisted of 129 recordings from 49 patients with class II obesity.

Preprocessing steps

Preprocessing steps were common for all datasets, and were performed in Python using custom and freely available scripts, including MNE-Python [32]. Recordings were bandpass filtered from 0.5 to 50 Hz and subsequently down-sampled to 100 Hz using the functions *filter_data* and *resample* from the MNE package. We then generated a bipolar montage consisting of four channels (F3-C3, C3-O1, F4-C4, and C4-O2) for each recording. These specific channels were chosen because they can be derived from both PSG EEG electrodes and long-term scalp EEG electrodes, regardless of the reference electrodes used for recording. This step renders the inputs "reference channel-free," that is, independent of the reference channel used during the recording. The EEG was segmented into non-overlapping 30 s epochs and assigned a sleep stage based on the original expert annotations. Artifactual epochs (defined as epochs in which at least 1 channel had an absolute value of amplitude > 500 μV or demonstrated a flat signal longer than 5 s) were removed. This led to the removal of ~10.5% of EEG segments. Most (94.21%) of these artifactual segments occurred during the awake state. All artifactual segments were excluded from use in both training and validation/testing procedures for our sleep staging algorithms. Recordings from all datasets were scaled so that each channel's median and interquartile range of voltages matched that of the MGH–PSG dataset.

Model architecture

The sleep staging model consists of two parts: a CNN followed by a RNN. Each 30 s window of scalp EEG data is input to the CNN, which learns a representation of the raw signal and acts as a feature extractor, bypassing the need for handcrafted features. Each CNN outputs a feature vector, which is a representation of the original 30 s EEG window in a reduced dimensional space. These feature vectors are then input to the RNN, which learns temporal dependencies present in the data and outputs sleep stages given a sequence of these feature vectors.

An overview of the algorithm flowchart and operation at each sleep epoch is shown in Figure 1. The input to the entire CNN/RNN neural network is a sequence of N_{seq} consecutive, non-overlapping windows of 30-s EEG. For each 30-s window, the preprocessed signal from each bipolar channel is independently fed to a CNN. Each channel is processed with a stack of N_c convolutional layers with f_i filters composed of s_i samples ($i \in [1, Nc]$). Each convolutional layer is followed by a maxpooling layer of size = 8 and stride = 1. The outputs of each stack are then concatenated and flattened. The resulting vector, which contains the feature representations from all four bipolar channels from a single 30-s window, is then passed through a series of N_R layers of bi-directional RNN cells. Each layer of the RNN is composed of N_{seq} RNN cells. The overall RNN structure thus consists of $N_{seq} \times N_R$ RNN cells. The type of RNN cell that we use is the long short term memory cell [33]. Each RNN cell is composed of N_{units} neurons. The final layer is a five-unit fully connected layer with a softmax output activation. Of note, N_c, f_i, s_i, N_R, N_{seq}, and N_{units} are hyperparameters that were tuned as described in the next subsection.

We used the rectified linear unit (relu) function [34] as the activation function in the convolutional layers, while the hard sigmoid and tanh functions were, respectively, used as the activation function for the gates and states of the RNN cell.

Model tuning and training

We used the Keras library [35] running on top of Tensorflow [36] to build and configure the CNN and RNN models. These were trained on two CUDA-enabled NVIDIA GPUs, running on CentOS 7. The MGH–PSG dataset was divided into training, validation, and testing sets containing 5,041, 650, and 650 different patients, respectively. Each model was trained separately using a two-step training procedure. We first trained the CNN alone, and then subsequently trained the combined CNN/RNN model.

Training the CNN alone

Initial training of the CNN utilized the MGH–PSG dataset. In the first step, the CNN was trained to extract useful time-invariant features from each EEG channel, resulting in a featurized representation of a 30 s EEG segment. This was achieved by isolating

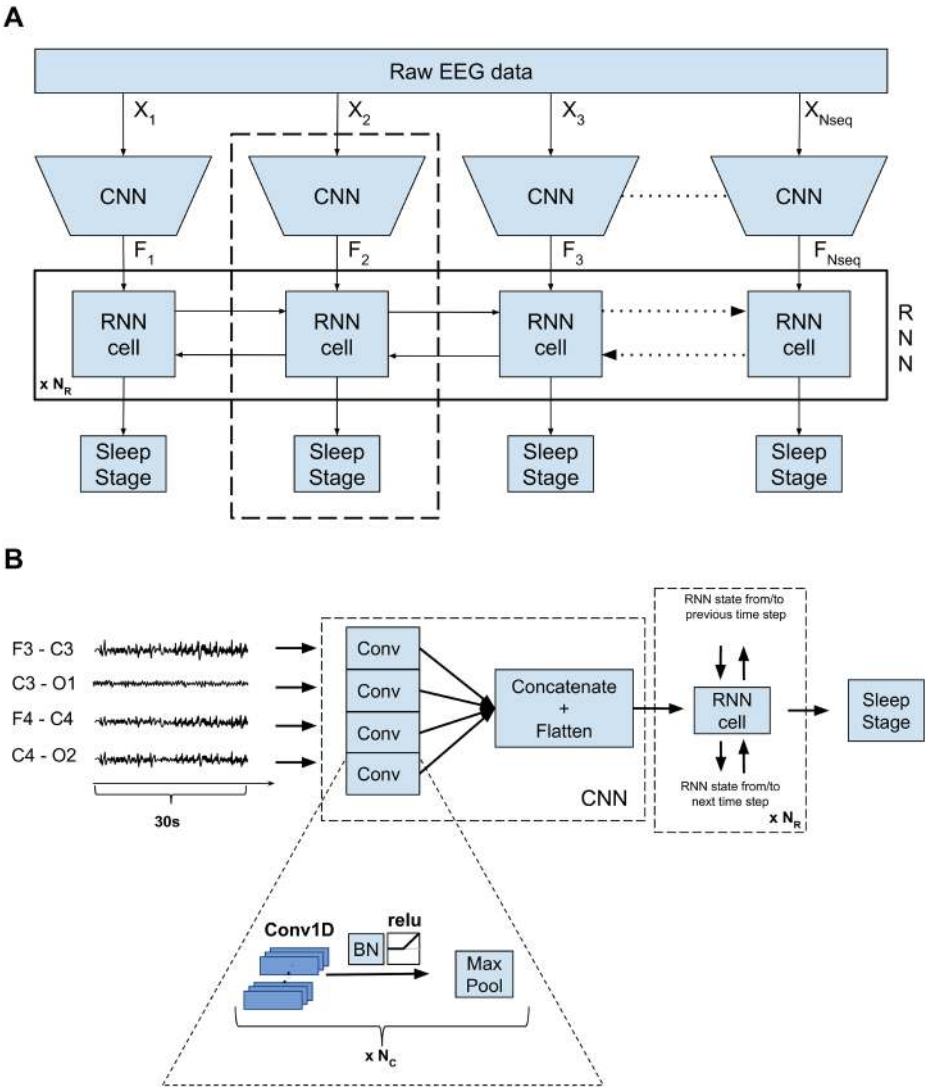

Figure 1. (A) Architecture of the deep neural network for sleep staging, from the perspective of a sequence of 30-s EEG window inputs. The input to the neural network CRNN is a sequence of N_{seq} consecutive 30-s EEG windows. Each 30-s window X_i is fed into a CNN, which outputs a featured representation of that segment F_i that goes into an individual cell of a RNN. Each RNN cell will then output a sleep stage corresponding to each element of the EEG sequence. The dashed box corresponds to the processing stream for one 30-s EEG window, and the architecture is further detailed in (B). (B) Architecture of the deep neural network for sleep staging, from the perspective of a single 30-s EEG window. For a given 30-s EEG window, each individual bipolar signal is independently fed into a stack of N_c convolutional layers (Conv). The outputs from each stack are then concatenated and flattened, resulting in a feature vector that is fed to a stack of RNN cells with N_R layers, which finally output the predicted sleep stage for this timestep. Further details regarding the architecture are provided in the text. The blue boxes under Conv1D represent the filters in the convolutional layer. BN, batch normalization; FC, fully connected layer.

the CNN from the entire model and appending it with a softmax layer to predict the sleep stage of the input epoch. The network was trained to recognize sleep stages from an input of size 4 × 3,000, representing a 30 s EEG segment with four bipolar channels, sampled at 100 Hz. Given the imbalance in sleep stage classes in the MGH–PSG dataset, we modified the cost function such that the penalties for sleep stage misclassifications were inversely proportional to class frequency. We used batch normalization [37] after each convolution and before every activation. The CNN was trained using the Adaptive Moment Estimated (Adam) optimization algorithm [38], with a batch size of 128, and parameters β_1, β_2, and ε set to 0.9, 0.999, and 10^{-8}, respectively. After completion of each training cycle (i.e. a complete iteration over all training examples), we calculated the performance of the CNN model on the validation data. The learning rate was reduced by half when the validation accuracy did not increase for three consecutive cycles and training was stopped when the validation accuracy did not increase for six consecutive cycles. We kept the model weights that were obtained at the cycle with the highest validation accuracy.

This procedure was repeated for different sets of randomly chosen values for N_c, f_i, s_i, and learning rates, resulting in the testing of different models. The hyperparameters and range of values evaluated for the CNN are shown in Supplementary Table S1.

The CNN model and hyperparameter set with the best validation accuracy and kappa was retained for the next part of the training procedure and will be referred to hereafter as CNN_{PSG}. CNN_{PSG} consisted of three convolutional layers with 8, 16, and 32 kernels each, trained using a learning rate of 0.001. The filters

of the first convolutional layer have a size of 50 samples, while those of the following 2 layers have a size of 8 samples.

Training the combined CNN/RNN model

In the second step, we trained the whole CNN/RNN model. The weights of the CNN were set as in CNN_{PSG} and were fixed throughout this second training procedure. From the training data, we generated training sequences of size N_{seq}, where each element of the sequence is a 4-channel, 30-s EEG window. Training sequences were generated as follows. For a certain recording, let X_i be the i-th non-overlapping, non-artifactual 30-s EEG window in that recording and N_W be the total number of non-artifactual 30-s windows. A training sequence S_M is composed of elements (Xi)i∈[M, M+Nseq-1], where M ∈[1, $N_W - N_{seq}$]. Therefore, for a certain record, up to $N_W - N_{seq}$ different training sequences can be extracted. Note that training sequences always remained the same length, but were not always contiguous in time, since artifactual segments were omitted prior to the formation of these sequences.

A training cycle was defined as a complete iteration over 16,384 unique training sequences, which were chosen randomly across all training patients at the beginning of each new training cycle. The RNN was trained using the Adam optimization algorithm [38], with a batch size of 128, and parameters β_1, β_2, and ε set to 0.9, 0.999, and 10^{-8}, respectively. After each training cycle, the performance of the model in training was evaluated on the validation data. The validation data was comprised of the original non-overlapping 30 s EEG segments that were scored by the sleep technician. Similar to the training procedure used for the CNN alone, the learning rate was divided in half if the validation accuracy did not improve for three cycles, and training was halted once it did not improve for six consecutive cycles. To prevent overfitting, we used dropout regularization [39] on each RNN layer's feedforward path [40], with a dropout probability of D_R. Note that for this part, no modifications to the cost function were made, that is, all classes were treated equally.

This training process was repeated for different sets of randomly chosen values for N_{RNN}, N_{units}, D_R, N_{seq}, and learning rates. The hyperparameters and range of values evaluated for the RNN are shown in Supplementary Table S1. The model with the best accuracy on the validation set was saved, and its performance was calculated on the testing data. We will refer to this model hereafter as $CRNN_{PSG}$. The RNN portion of $CRNN_{PSG}$ consisted of 4 RNN layers with 64 units each, trained with a dropout rate of 0.5. The optimal input sequence length to $CRNN_{PSG}$ (N_{seq}) was 32.

Transfer learning

After developing $CRNN_{PSG}$ on the MGH–PSG dataset, we then fine-tuned this algorithm specifically for use on long-term scalp EEG data, using transfer learning. For this part, rather than using the PSG dataset, we used the Ambulatory scalpEEG dataset to generate the training and validation data. We used the same CNN/RNN network architecture as developed above, and used the optimal weights determined for $CRNN_{PSG}$ to initialize the weights for transfer learning. We fixed the weights for the CNN portion, and re-trained the model to fine-tune the weights for the RNN portion. The final model developed using this transfer learning procedure is referred to as $CRNN_{EEG}$.

Since the model architecture was fixed, a random search was carried out to determine the optimal set of hyperparameters for the training procedure, namely the learning rate, batch size, number of training cycles, and balancing scheme. The different hyperparameters and range of values that were considered are shown in Supplementary Table S2.

A training cycle was defined as a complete iteration over N_x/N_{seq} sequences, where N_x is the total number of non-overlapping 30 s segments in the training set. We randomly generated a new set of N_x/N_{seq} training sequences across all training patients at the start of each new training cycle, similar to as described in the training procedure for $CRNN_{PSG}$.

We used fivefold cross validation to carry out our search for the optimal hyperparameters. Patients were partitioned into five folds. For a given hyperparameter set, the neural network was trained on data from four folds and tested on the data from the left-out fold (testing fold). This was repeated five times, with each fold being used once as the testing fold. The test fold results from each iteration were aggregated across all folds to determine the performance for this hyperparameter set. We also retained the model's performance on individual patients in each test fold. This procedure was repeated for each set in the hyperparameters grid, to find the one with the best performance. The RNN portion was trained using the Adam optimization algorithm, with parameters as above.

The optimal performance for $CRNN_{EEG}$ was achieved with a learning rate of 0.00077, batch size of 1,024, 10 training cycles, and penalizing the cost function according to class frequencies. $CRNN_{EEG}$ is publicly available and can be found on https://github.com/mauriceaj/CRNNeeg-sleep.

Performance evaluation

To measure performance on sleep stage classification, we report the percentage of agreement (accuracy) and the Cohen's kappa statistic [41]. Cohen's kappa takes into account the probability of agreement occurring by chance and is a metric commonly used for sleep staging performance. We primarily report an overall Cohen's kappa (pooled performance across all sleep epochs and all recordings within a dataset).

To determine optimal hyperparameters for model selection, we used a single metric, which we chose as the average between the overall Cohen's kappa and the class-averaged agreement. The class-averaged agreement was calculated by taking the class-specific agreement within each of the five sleep stages and averaging these values. The class-specific agreement within Stage S is the percentage of all stages S that were correctly identified by the sleep staging algorithm. We also computed confusion matrices, in which each element (i, j) represents the empirical probability of predicting class j given that the expert label is class i.

For the MGH–PSG dataset, we report the performance of $CRNN_{PSG}$ on the independent holdout test set. For the ambulatory scalpEEG dataset, we report the cross-validated performance of $CRNN_{EEG}$, averaged across the five folds.

Benchmark comparison

For benchmark comparison to a previously published deep learning algorithm for sleep staging, we implemented

SeqSleepNet, a hierarchical RNN developed by Phan et al. [42, 43]. In SeqSleepNet, each 30 s EEG signal goes through a filterbank layer for preprocessing and then an attention-based recurrent layer for short-term sequential modelling. A sequence of such epochs are then processed by a recurrent layer that models long-term time-dependencies of sleep stages. Similar to our strategy, the algorithm is first trained on a larger dataset (source domain, Montreal Archive of Sleep Studies [MASS]) [44]. The algorithm is then fine-tuned and tested on a smaller dataset (target domain). A key difference between our proposed method and theirs, is that the input channels for SeqSleepNet are different between the source and target domains.

Here, we started with the publicly available model of SeqSleepNet that was pre-trained on the MASS dataset using the C4-A1 channel (https://github.com/pquochuy/sleep_transfer_learning). Using the scalpEEG dataset, we performed transfer learning with fivefold cross-validation, to fine-tune SeqSleepNet to use the F4-C4 channel for sleep staging. For transfer learning, we used the same training parameters as described previously and that were provided with the algorithm [42, 43].

Statistical analysis

Population statistics and performance results are reported as mean ± standard deviation unless stated otherwise. Correlation between two variables was assessed using the Pearson correlation coefficient. Differences between two groups were assessed using an independent samples *t*-test.

Results

$CRNN_{PSG}$ performs at human expert level on the MGH–PSG dataset

On the MGH–PSG holdout test dataset, $CRNN_{PSG}$ achieved an overall agreement of 81.1%, a class-averaged agreement of 78.2%, and an overall Cohen's kappa of 0.74, comparable to the reported Cohen's kappa for human expert level inter-rater agreement in sleep staging (0.75 ± 0.11) [45, 46]. $CRNN_{PSG}$ correctly classified Wake, N1, N2, N3, and REM stages 82, 54, 87, 78, and 90% of the time, respectively, as shown in Supplementary Figure S1, A.

Transfer learning on $CRNN_{PSG}$ optimizes sleep staging performance on long-term scalp EEG data

Applying $CRNN_{PSG}$ to the ambulatory scalpEEG dataset without any fine tuning resulted in an overall agreement of 83.3%, a class-averaged agreement of 71.6%, and a Cohen's kappa of 0.68. The resulting confusion matrix is shown in Supplementary Figure S1, B. Applying only the CNN portion of the algorithm to the ambulatory scalpEEG dataset resulted in an overall agreement of 75.8%, a class-averaged agreement of 67.0%, and a Cohen's kappa of 0.56, demonstrating that addition of the RNN substantially improved the algorithm's performance.

To optimize the performance of $CRNN_{PSG}$ specifically for long-term scalp EEG studies, we used transfer learning with fivefold cross-validation. The re-optimized algorithm, $CRNN_{EEG}$, resulted in improved performance on the ambulatory scalpEEG dataset, with a cross-validated agreement of 88.9 ± 1.3%, class-averaged agreement of 74.1 ± 0.02%, and Cohen's kappa of 0.78 ± 0.02 (Table 2). Figure 2, A shows a representative hypnogram of a patient's long-term scalp EEG data over 24 h, as staged by the sleep technician and by $CRNN_{EEG}$. The confusion matrix for $CRNN_{EEG}$ is shown in Figure 2, B, showing the highest agreement for awake stages (94%) and the lowest agreement for stage N1 (53%).

Grouping stages N1, N2, and N3 into a single NREM class resulted in a cross-validated overall agreement of 92.6 ± 0.8% and a Cohen's kappa of 0.85 ± 0.01. NREM sleep was correctly classified 92% of the time, as shown in Figure 2, C.

We next calculated $CRNN_{EEG}$'s performance for sleep staging at night, compared to during the day. Between 8:00 pm and 8:00 am, $CRNN_{EEG}$'s overall agreement and kappa were 82.8 ±1.8% and 0.74 ± 0.02, respectively. Between 8:00 am and 8:00 pm, $CRNN_{EEG}$'s overall agreement and kappa were 96.3 ± 1.0 and 0.77 ± 0.01, respectively. The confusion matrices for both time periods are shown in Supplementary Figure S2. This indicates that $CRNN_{EEG}$ performs well at staging sleep both during the day (e.g. daytime naps), as well as at night.

Evaluating $CRNN_{EEG}$ performance during transition and stable sleep periods

Sleep staging accuracy and inter-rater agreement are generally worse during transition periods between sleep stages [47]. As such, we evaluated $CRNN_{EEG}$'s performance during transition segments, where a 30 s EEG segment of label L was defined as a transition epoch if any of the p previous or p following 30 s segments had a label different than L. Conversely, a 30 s EEG segment of label L was defined as a stable epoch if all the p previous and p following 30 s segments also had the label L. The cross-validated accuracy for $CRNN_{EEG}$ on transition epochs ranged from 51.5% and 64.3% depending on p, while Cohen's kappa ranged from 0.33 and 0.51 (Supplementary Figure S3, A). In contrast, $CRNN_{EEG}$'s performance on stable epochs was much higher, with accuracy ranging from 92.9% to 96.5%, and Cohen's kappa ranging from 0.84 to 0.90, depending on p (Supplementary Figure S3, A). The confusion matrix for stable epochs at $p = 1$ is shown in Supplementary Figure S3, B.

The above result demonstrates that one way to ensure the accuracy of $CRNN_{EEG}$'s output is to consider only stable epochs. However, the definitions of stable and transition epochs as above were based on the "ground truth" (expert rated) data, whereas in real-world applications of $CRNN_{EEG}$, this ground truth information on sleep state stability would not be known.

Table 2. Performance of $CRNN_{EEG}$ and $CRNN_{PSG}$ across studied datasets

	MGH–PSG		ScalpEEG		Home PAP		ABC	
	Kappa	Accuracy (%)	Kappa	Accuracy (%)	Kappa	Accuracy (%)	Kappa	Accuracy (%)
$CRNN_{PSG}$	0.74	81.1	0.68	83.3	0.64	73.7	0.64	75.1
$CRNN_{EEG}$	0.67	75.8	0.78	88.9	0.69	77.3	0.66	74.9

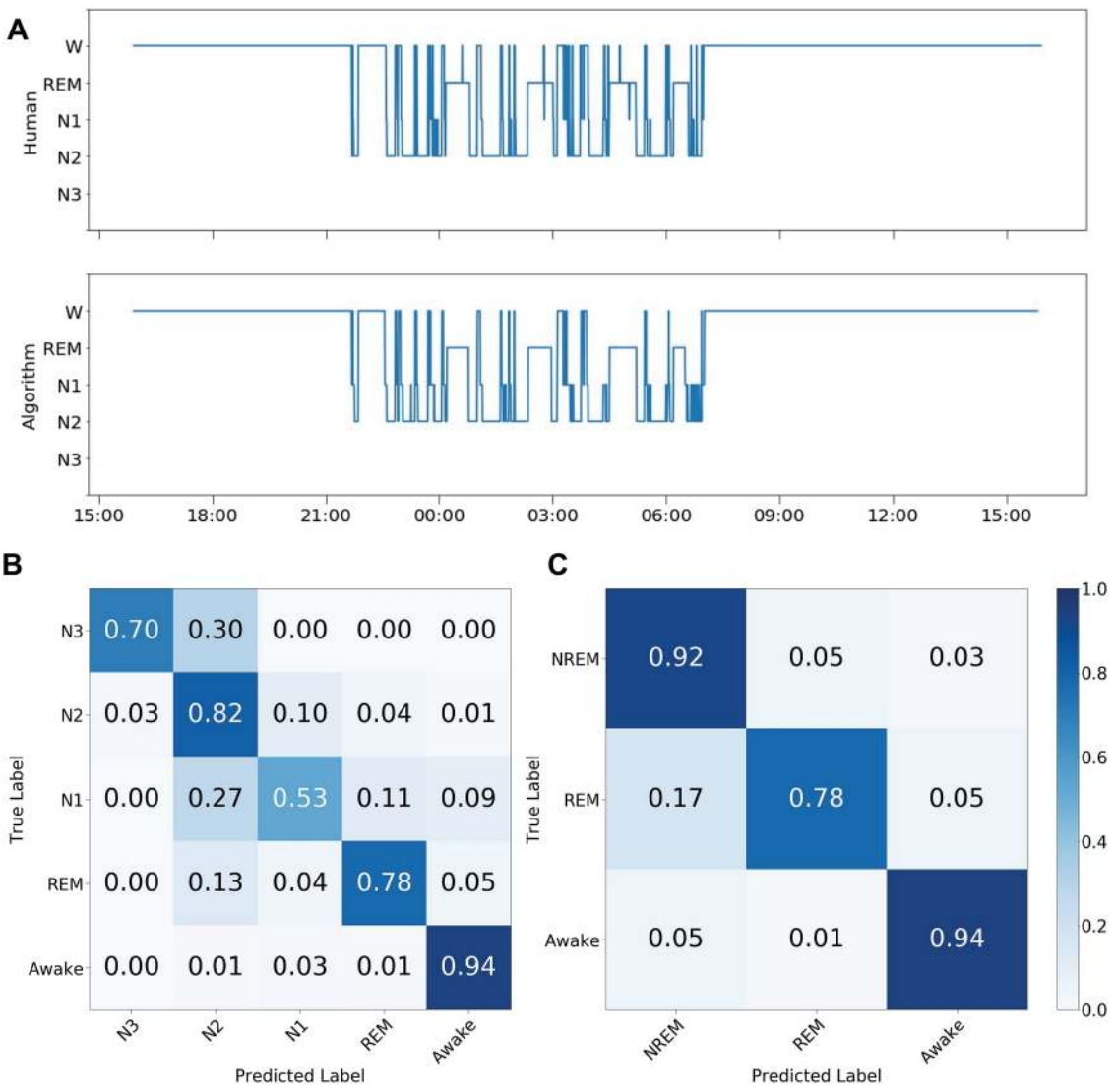

Figure 2. Performance of $CRNN_{EEG}$ on scalpEEG dataset. (A) Example output hypnogram of a 24 h ambulatory EEG scored by a sleep technician and by $CRNN_{EEG}$. (B) Confusion Matrices for $CRNN_{EEG}$ on the scalpEEG dataset for (B) all sleep stages and (C) NREM/REM/wake stages.

Therefore, we next sought to characterize $CRNN_{EEG}$'s performance with respect to the stability of its own output. We used the same definitions for stable and transition epochs above, but now used $CRNN_{EEG}$'s output, rather than the ground truth data, to define these epochs. Using these algorithm output-based definitions, we found that for transition epochs, the cross-validated accuracy and Cohen's kappa ranged from 52.8% to 63.5% and 0.36 to 0.50, respectively. For stable epochs, the overall cross-validated accuracy and Cohen's kappa ranged from 92.0% to 96.2%, and 0.83 to 0.89, respectively (Figure 3, A). We also evaluated $CRNN_{EEG}$'s performance on individual sleep stages based on transition versus stable epochs (Figure 3, B and C). From a practical standpoint, if one were interested primarily in identifying N2 sleep epochs from a long-term scalp EEG recording with high accuracy, this figure demonstrates that choosing epochs with an algorithm output stability defined by a window of size $p = 2$ would result in an accuracy of 88.7%, with the tradeoff of discarding 25.9% of all epochs labeled as N2. Ultimately, these results suggest that high confidence can be given to stable outputs of $CRNN_{EEG}$, whereas transition outputs should be considered with some caution.

$CRNN_{EEG}$ performance is robust to age and common EEG background abnormalities

Ambulatory EEGs recordings are often performed on patients with a wide range of ages, and whose EEGs may have background abnormalities such as focal or generalized slowing, or epileptiform discharges.

Among healthy control subjects in our ambulatory scalpEEG dataset, we found no correlation between $CRNN_{EEG}$ performance and age (Figure 4, A). The correlation coefficient for Cohen's kappa and age was –0.169 ($p = 0.218$), and the correlation coefficient for overall accuracy and age was –0.109 ($p = 0.410$). These results show that $CRNN_{EEG}$ can be applied to a wide range of ages across adults.

To determine how different disease states and EEG background abnormalities might affect $CRNN_{EEG}$'s performance, we evaluated its performance on sub-groups of our ambulatory

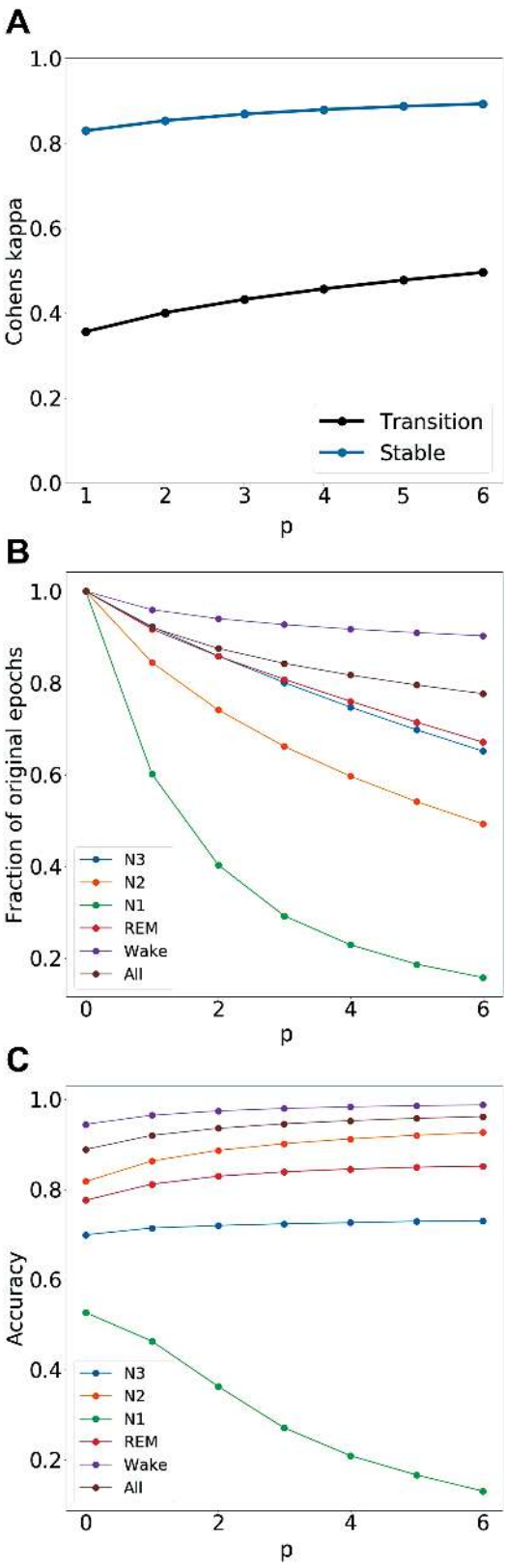

Figure 3. Performance of $CRNN_{EEG}$ on stable and transition epochs, as defined by the output of CRNNEEG. (A) Cohen's kappa for $CRNN_{EEG}$ on the scalpEEG dataset for transition and stable outputs, as a function of *p*. $CRNN_{EEG}$'s output is defined as a transition output if any of the previous or the following *p* predictions has a different sleep stage than that output. (B) Fraction of original epochs that are stable for each sleep stage, as a function of *p*, (C) accuracy of $CRNN_{EEG}$ on the stable outputs for each sleep stage, as a function of *p*.

scalpEEG dataset. Patients were divided into four groups: (1) cognitively normal (Normal); (2) probable Alzheimer's disease, with no history of epilepsy (AD); (3) probable Alzheimer's disease with epilepsy related to Alzheimer's (AD + Epilepsy); and (4) late-onset temporal lobe epilepsy of unclear etiology (Late onset epilepsy). Sub-group results were obtained by averaging the Cohen's kappa values (calculated during the cross-validation procedure) across individuals in each sub-group.

There was no difference in $CRNN_{EEG}$ performance across the different disease sub-groups (Figure 4, B). Moreover, $CRNN_{EEG}$'s performance was stable across a range of background abnormalities, including diffuse slowing, focal slowing, and spikes (epileptiform discharges) (Figure 4, B).

$CRNN_{EEG}$ generalizes well on novel datasets

We tested $CRNN_{EEG}$'s performance on two external datasets (homePAP and ABC), made available through the National Sleep Research Resource [28, 29]. Direct application of $CRNN_{EEG}$ to the homePAP and ABC datasets resulted in an overall Cohen's kappa of 0.69 and 0.66, respectively, with agreements of 77.3% and 74.6%. This corresponds to a substantial agreement [48] between $CRNN_{EEG}$ and human expert labels for both datasets. Performance on both datasets was within the range of human expert inter-rater reliability for sleep staging (0.75 ± 0.11) [45]. Thus, $CRNN_{EEG}$'s performance is generalizable across novel datasets.

$CRNN_{EEG}$ performance on patients with sleep-related disorders

We next calculated the performance of $CRNN_{EEG}$ on sub-groups of the MGH–PSG, homePAP, and ABC datasets with varying severity of obstructive sleep apnea (OSA). For each recording, we extracted the Apnea–Hypopnea Index (AHI), a measure of OSA severity, and classified the recording as showing: no OSA (AHI < 5); mild to moderate OSA (5 ≤ AHI ≤ 30); and severe OSA (AHI > 30). The results for each sub-group are shown in Table 3. For all three datasets, the algorithm performed significantly worse on patients with severe OSA compared to those with non-severe OSA (including no OSA, and mild to moderate OSA). This is not surprising, as patients with severe OSA have frequent arousals/awakenings during sleep, and thus more transition epochs. Consistent with this, we found that the percentage of stage N1 sleep increased with OSA severity in all three datasets (Table 3). Notably, while $CRNN_{EEG}$ applied to the HomePAP dataset overall had a kappa of 0.69, its overall performance on participants with non-severe OSA was 0.71, and on participants with no OSA was 0.74. Similarly, while $CRNN_{EEG}$ applied to the ABC dataset overall had a kappa of 0.66, its performance on participants with non-severe OSA was 0.71, and on participants with no OSA was 0.71 (Table 3). Thus, while $CRNN_{EEG}$ performed overall well on the external PSG datasets, its performance on these datasets was significantly better when tested on patients without severe OSA.

We also calculated the performance of $CRNN_{EEG}$ on a subgroup of patients with chronic insomnia. These data were collected as part of the Bioboosti study (a subset of the scalpEEG dataset), which consisted of 29 recordings of 19 patients with complaints of chronic insomnia. For the Bioboosti dataset, $CRNN_{EEG}$ achieved a Cohen's kappa of 0.79 ± 0.08, indicating excellent performance on patients with insomnia.

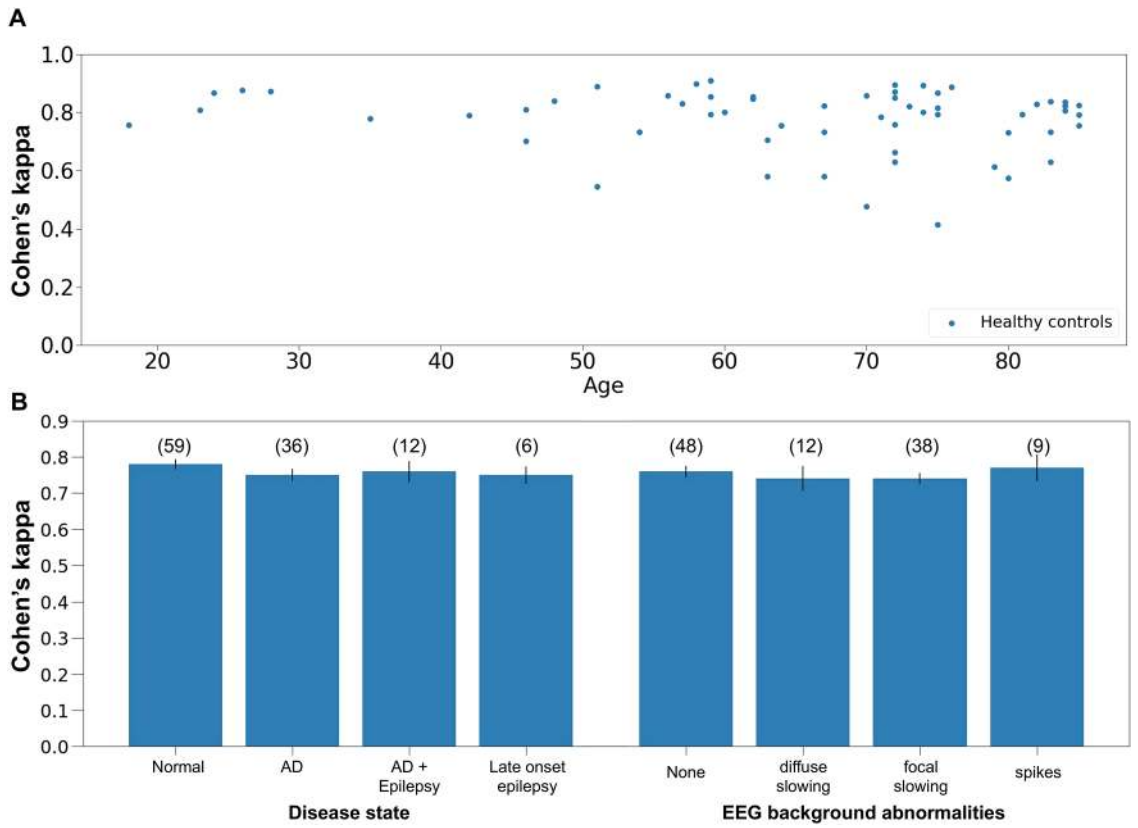

Figure 4. Performance of $CRNN_{EEG}$ across different ages, disease states, and EEG background abnormalities. (A) Cohen's kappa for $CRNN_{EEG}$ on the scalpEEG dataset as a function of age for cognitively normal subjects. (B) Averaged Cohen's kappa for $CRNN_{EEG}$ across patients with disease states and with different EEG background abnormalities.

Table 3. Performance of $CRNN_{EEG}$ on patients with OSA

Dataset	No OSA	Mild to moderate OSA	Severe OSA
Number of recordings (% of dataset)			
MGH–PSG†	356 (45.0)	380 (48.1)	55 (6.9)
HomePAP	74 (30.4)	103 (42.4)	66 (27.2)
ABC	2 (1.5)	45 (34.9)	82 (63.6)
N1 sleep (% total recording time)‡			
MGH–PSG†	10.7	15.3	18.0*
HomePAP	8.5	10.3	14.4**
ABC	7.6	10.1	17.0**
Overall Cohen's Kappa‡			
MGH–PSG†	0.68	0.67	0.61 **
HomePAP	0.74	0.68	0.65 **
ABC	0.71	0.70	0.63 **

†Refers to the holdout testing set from MGH–PSG (includes 791 recordings from 650 patients).
‡Statistically significant differences between severe and non-severe OSA groups are shown as follows: *$p < 0.01$ and **$p < 0.001$. The non-severe OSA group combines the No OSA and the Mild to Moderate OSA groups.

Evaluation of different transfer learning strategies

In our transfer learning method described above for $CRNN_{EEG}$, we kept the CNN weights fixed and fine-tuned the RNN weights. We also experimented with fine-tuning both the CNN and RNN weights during transfer learning. This resulted in a cross-validated Cohen's kappa of 0.78 ± 0.02 and an overall agreement of 88.9 ± 1.1%, performance that was not significantly different from fine-tuning the RNN weights alone. This likely indicates that the features learned by pretraining the CNN on the large MGH–PSG dataset are generalizable to other datasets.

To examine the utility of our transfer learning approach, we also experimented with training our model from scratch using the ambulatory scalpEEG dataset (without first pre-training it on the MGH–PSG dataset). This resulted in a cross-validated Cohen's kappa of 0.76 ± 0.03 and agreement of 88.1 ± 1.5%, comparable to the performance of $CRNN_{EEG}$ on the scalpEEG dataset. However, applying this new model to the homePAP and ABC datasets resulted in a Cohen's kappa of only 0.59 and 0.55, respectively, with overall agreements of 70.9% and 67.0%. Training the deep learning model from scratch using only the scalpEEG dataset clearly led to overfitting. This supports our initial strategy of learning feature extraction on a large dataset first, followed by transfer learning and fine-tuning on a smaller, more relevant dataset.

Comparison of $CRNN_{EEG}$ to another deep learning algorithm for sleep staging

We performed a direct comparison of $CRNN_{EEG}$ to SeqSleepNet, a previously published deep learning algorithm for sleep staging that also used a transfer learning approach. SeqSleepNet reported an accuracy of 85.5% and a Cohen's kappa of 0.79 on its testing dataset [42, 43].

We started with a pre-trained version of SeqSleepNet, and using transfer learning on the scalpEEG dataset, we fine-tuned all model weights to use the F4-C4 channel for sleep staging. The

fine-tuned SeqSleepNet algorithm achieved a cross-validated Cohen's kappa of 0.76 ± 0.01 and an overall agreement of 89.5 ± 0.1%. These results were overall similar to the cross-validation performance of $CRNN_{EEG}$ on the scalpEEG dataset (Cohen's kappa of 0.78 ± 0.02 and overall agreement of 88.9 ± 1.3%). However, we found that the performance of the fine-tuned SeqSleepNet did not generalize as well to other datasets as $CRNN_{EEG}$ did. Application of the fine-tuned SeqSleepNet to the HomePAP dataset yielded a Cohen's kappa of 0.55 and an overall accuracy of 67.7%. On ABC, the fine-tuned SeqSleepNet achieved a Cohen's kappa of 0.43 and an overall accuracy of 59.0%.

Discussion

This work describes the development of an automated sleep staging algorithm ($CRNN_{EEG}$) that can be used specifically for long-term (> 12 h) scalp EEG recordings. Our algorithm's performance is comparable to that of expert level inter-rater agreement for sleep staging. A prior study of 72 records from three European sleep laboratories found that pairwise sleep expert inter-rater agreement has an average Cohen's kappa of 0.75 ± 0.11 [45], which is consistent with the reported Cohen's kappa of 0.75 ± 0.01 between American and Chinese sleep centers for 40 subjects [46]. Our algorithm initially achieved a Cohen's kappa of 0.74 on PSG data, and when fine-tuned on ambulatory scalp EEG data, achieved a Cohen's kappa of 0.78. Importantly, on validation with two novel datasets, HomePAP and ABC, our algorithm achieved a Cohen's kappa of 0.66 and 0.69, respectively. When patients with severe OSA were excluded from those datasets, our algorithm achieved a Cohen's kappa of 0.71 on both datasets.

While large clinical PSG datasets have allowed the development of powerful algorithms for automated sleep staging of PSGs, these methods have not been easily translatable for use on long-term scalp EEG data, largely due to conventional differences between recording PSGs and recording long-term EEGs. Namely, not all physiologic measures captured on PSG are typically available with long-term EEG, and moreover, the convention of using a contralateral mastoid reference for PSG EEG channels is rarely used in long-term scalp EEGs. Development of algorithms for automated sleep staging of long-term scalp EEG data has moreover been hampered by the fact that these clinical studies do not typically include manual sleep staging of the data. Consequently, there is a paucity of clinical ambulatory scalp EEG data with labeled sleep stages to use for training such algorithms.

Here, to leverage the "big data" power of a large, expert-labeled clinical PSG dataset, yet still allow translation to long-term scalp EEG recordings, we first derived reference channel-free scalp EEG bipolar channels from the PSG data. Notably, these scalp EEG channels can easily be derived from any International 10–20 System scalp EEG recording, regardless of the reference electrode used. We then applied deep learning methods to the large PSG dataset to train an automated sleep staging algorithm that uses these reference channel-free scalp EEG channels as the inputs. We first trained a CNN to extract the relevant features from these inputs. To further improve performance, we added an RNN that integrated the CNN-extracted features along with the temporal dependencies of the data. Finally, application of transfer learning methods allowed us to tailor the algorithm specifically for use on long-term scalp EEG data.

Notably, for transfer learning, we kept the CNN weights fixed and only fine-tuned the weights of the RNN. The success of this approach implies that the featurized representation of sleep stages learned by the CNN is generalizable and transferable across datasets, as we only needed to fine-tune the decision rules to the temporal dependencies that were specific to the scalpEEG dataset. This is further reinforced by the fact that additional fine-tuning of the CNN portion of the algorithm did not further improve performance. Long-term scalp EEG datasets differ largely in sleep stage content compared to PSG datasets, as long-term scalp EEGs contain a much greater proportion of the awake state compared to PSGs. We hypothesize that the temporal dependency of these sleep class differences was a large factor in what was optimized in the transfer learning process.

On benchmark comparison to SeqSleepNet, a previously published deep learning algorithm for automated sleep staging, we found that our algorithm's performance was superior, particularly as it demonstrated better generalizability on novel datasets. Notably, $CRNN_{EEG}$ was pre-trained on the MGH–PSG dataset with over 6,400 patients, whereas SeqSleepNet was pre-trained on the MASS dataset with 200 patients. It is likely that training on data from only 200 patients results in model overfitting. This is consistent with what we found when we attempted to train $CRNN_{EEG}$ from scratch using the scalpEEG dataset (113 patients each with 24–72 h of data, thus likely similar in size to the MASS dataset). Namely, when trained only on scalpEEG, $CRNN_{EEG}$'s cross-validation performance was excellent, but its generalizability to external datasets was poor. This underscores the importance of training deep learning algorithms on sufficiently large datasets to prevent overfitting.

Our algorithm has a number of important applications. First, it can provide information on sleep macro-architecture and sleep quality. Second, it allows isolation of specific sleep stages, for example, to analyze sleep microarchitecture or other features related to that particular sleep stage, such as sleep spindle analysis or slow-wave spindle coupling during NREM sleep. Third, it provides a powerful tool for large-scale analyses of the relationships between sleep and epileptiform abnormalities, especially when combined with other automated methods for spike and seizure detection [49]. Seizures may occur preferentially from different sleep stages [50], and REM-related interictal discharges may have a particular importance in epilepsy localization [51]. Automated sleep scoring paired with seizure or spike detection could be particularly useful for chronotherapy in epilepsy, where medications are specifically given at times when a particular patient is at greatest risk for having seizures, or when the greatest amount of epileptiform activity is noted to occur [52]. Automated sleep scoring of EEG recordings may also be of major help to clinicians and researchers in the epilepsy and neurophysiology community. Healthy sleep is an increasingly recognized necessity for multiple diseases, including epilepsy [53]. Sleep evaluations are also important for an array of other disorders, and it is impracticable and expensive to perform separate tests in these patients and ignore the vast sleep-related EEG information acquired during an EMU study or ambulatory outpatient EEG monitoring.

Our study has several limitations. First, sleep scoring on the scalpEEG dataset was performed by a single sleep technologist, and not all scalp EEG recordings had EOG and EMG channels available for sleep staging. While it would have been optimal to have a ground truth sleep-scored dataset based on the consensus of

multiple expert sleep raters, in reality, the vast majority of clinical PSGs are scored by a single sleep technologist, with diagnostic and treatment decisions made on the basis of this. Regardless, testing $CRNN_{EEG}$ on two novel datasets showed substantial agreement with human expert labeling, indicating that our model is generalizable to unseen datasets. Second, while there was a wide range of ages represented in both datasets, our patient population was skewed towards older adults. Nevertheless, the performance of $CRNN_{EEG}$ was comparable on both young and older adults, and there was no correlation between algorithm performance and age. It is unclear how $CRNN_{EEG}$ would perform in a pediatric population. Third, $CRNN_{EEG}$ has primarily been tested on patients with relatively normal EEG backgrounds. Therefore, it is unclear how the algorithm would perform in patients with grossly abnormal EEGs, for example, critically ill patients on sedative medications, or patients with significant brain injuries, congenital brain malformations, or other neurologic diseases resulting in more striking EEG background abnormalities. However, our algorithm was relatively robust to more common EEG background abnormalities, including generalized and focal slowing, and sporadic epileptiform discharges. Fourth, $CRNN_{EEG}$ performs worse on patients with severe OSA compared to patients with non-severe OSA—albeit still with moderate to substantial agreement with expert labeling. One contributing factor to this difference is the higher proportion of stage N1 found in patients with severe OSA, which is more difficult to stage by algorithms and experts alike. Last, we developed $CRNN_{EEG}$ using only raw EEG signals as inputs. Recent studies have shown that deep neural networks that use spectrogram representations of EEG segments lead to better classification accuracy compared to those that use raw EEG segments [25, 26]. As such, it is possible that addition of spectrogram representations to our algorithm could result in further improved performance.

In conclusion, this work presents an important new tool for automated sleep staging of long-term scalp EEG data, that performs at the level of human expert sleep scorers. In addition to potentially expanding access to sleep-related diagnostics, it will facilitate the use of large clinical long-term EEG datasets for sleep research purposes. $CRNN_{EEG}$ is publicly available and can be found on https://github.com/mauriceaj/CRNNeeg-sleep.

Supplementary material

Supplementary material is available at *SLEEP* online.

Funding

We thank Jonathan Pham for technical assistance. ADL was funded by grants from the National Insitute of Health-National Institute of Neurological Disorders and Stroke (NIH-NINDS) K23 NS01037, and the American Academy of Neurology Institute. MBW was funded by a Breakthroughs in Gerontology Grant, supported by the Glenn Foundation for Medical Research and the American Federation for Aging Research, an AASM Foundation Strategic Research Award, and by grants from the National Institutes of Health (1R01NS102190, 1R01NS102574, 1R01NS107291, and 1RF1AG064312). SSC was funded by grants from the NIH-NINDS R01 NS062092 and K24 NS088568. RAS was funded by the AJ Trustey Epilepsy Research Endowed Fund. The NSRR is supported by Grant Number HL114473 from the National Heart, Lung, and Blood Institute, NIH.

Disclosure Statement

Financial Disclosure: none.

Non-financial Disclosure: none

References

1. Guld AT, *et al*. Drug taper during long-term video-EEG monitoring: efficiency and safety. *Acta Neurol Scand*. 2017;**135**(3):302–307.
2. Tolchin B, *et al*. Diagnostic yield of ambulatory EEGs in the elderly. *Clin Neurophysiol*. 2017;**128**(7):1350–1353.
3. Faulkner HJ, *et al*. The utility of prolonged outpatient ambulatory EEG. *Seizure*. 2012;**21**(7):491–495.
4. Punjabi NM, *et al*. Computer-assisted automated scoring of polysomnograms using the somnolyzer system. *Sleep*. 2015;**38**(10):1555–1566.
5. Malhotra A, *et al*. Performance of an automated polysomnography scoring system versus computer-assisted manual scoring. *Sleep*. 2013;**36**(4):573–582.
6. Younes M, *et al*. Accuracy of automatic polysomnography scoring using frontal electrodes. *J Clin Sleep Med*. 2016;**12**(05):735–746.
7. Mikkelsen KB, *et al*. Machine-learning-derived sleep–wake staging from around-the-ear electroencephalogram outperforms manual scoring and actigraphy. *J Sleep Res*. 2019;**28**(2):e12786.
8. Al-hussaini I, Xiao C, Westover MB, Sun J. SLEEPER : interpretable Sleep staging via Prototypes from Expert Rules. In: *Proceedings of the 4th Machine Learning for Healthcare Conference*. PMLR; 2019:721–739.
9. Virkkala J, *et al*. Automatic sleep stage classification using two-channel electro-oculography. *J Neurosci Methods*. 2007;**166**(1):109–115.
10. Fraiwan L, *et al*. Classification of sleep stages using multi-wavelet time frequency entropy and LDA. *Methods Inf Med*. 2010;**49**(3):230–237.
11. Liang SF, *et al*. A rule-based automatic sleep staging method. *J Neurosci Methods*. 2012;**205**(1):169–176.
12. Berthomier C, *et al*. Automatic analysis of single-channel sleep EEG: validation in healthy individuals. *Sleep*. 2007;**30**(11):1587–1595.
13. Wang Y, *et al*. Evaluation of an automated single-channel sleep staging algorithm. *Nat Sci Sleep*. 2015;**7**:101–111.
14. 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.
15. Hassan AR, Bhuiyan MIH. An automated method for sleep staging from EEG signals using normal inverse Gaussian parameters and adaptive boosting. *Neurocomputing*. 2017;**219**:76–87. doi:10.1016/j.neucom.2016.09.011
16. Khalighi S, *et al*. Automatic sleep staging: a computer assisted approach for optimal combination of features and polysomnographic channels. *Expert Syst Appl*. 2013;**40**(17):7046–7059.
17. 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.
18. 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.
19. Fraiwan LA, Khaswaneh NY, Lweesy KY. Automatic sleep stage scoring with wavelet packets based on single EEG recording. *World Acad Sci Eng Technol*. 2009;**3**(6):85–88.
20. 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.
21. Dong J, et al. Automated sleep staging technique based on the empirical mode decomposition algorithm: a preliminary study. *Adv Adapt Data Anal*. 2010;**02**(02):267–276.
22. Hsu YL, Yang YT, Wang JS, Hsu CY. Automatic sleep stage recurrent neural classifier using energy features of EEG signals. *Neurocomputing*. 2013. doi:10.1016/j.neucom.2012.11.003
23. Sun C, *et al*. Revisiting unreasonable effectiveness of data in deep learning era. *Proc IEEE Int Conf Comput Vis*. 2017;843–852. doi:10.1109/ICCV.2017.97
24. Bresch E, *et al*. Recurrent deep neural networks for real-time sleep stage classification from single channel EEG. *Front Comput Neurosci*. 2018;**12**:85.
25. Biswal S, *et al*. Expert-level sleep scoring with deep neural networks. *J Am Med Inform Assoc*. 2018;**25**(12):1643–1650.
26. 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
27. Pavlova MK, *et al*. Novel non-pharmacological insomnia treatment—a pilot study. *Nat Sci Sleep*. 2019;**11**:189–195. doi:10.2147/NSS.S198944
28. Dean DA 2nd, *et al*. Scaling up scientific discovery in sleep medicine: the national sleep research resource. *Sleep*. 2016;**39**(5):1151–1164.
29. Zhang GQ, *et al*. The national sleep research resource: towards a sleep data commons. *J Am Med Inform Assoc*. 2018;**25**(10):1351–1358.
30. Rosen CL, *et al*. A multisite randomized trial of portable sleep studies and positive airway pressure autotitration versus laboratory-based polysomnography for the diagnosis and treatment of obstructive sleep apnea: the HomePAP study. *Sleep*. 2012;**35**(6):757–767.
31. Bakker JP, *et al*. Gastric banding surgery versus continuous positive airway pressure for obstructive sleep apnea: a randomized controlled trial. *Am J Respir Crit Care Med*. 2018;**197**(8):1080–1083.
32. Gramfort A, *et al*. MEG and EEG data analysis with MNE-Python. *Front Neurosci*. 2013;**7**:267.
33. Hochreiter S, *et al*. Long short-term memory. *Neural Comput*. 1997;**9**(8):1735–1780.
34. Hinton N and, Nair V, Hinton GE. Rectified linear units improve restricted Boltzmann machines. In: proceedings of the 27th International Conference on Machine Learning. 2010:807–814. doi:10.1111/j.1365-2966.2012.21196.x
35. Chollet F. Keras: The Python Deep Learning library. 2015. https://keras.io/
36. Martín A, *et al*. TensorFlow: a system for large-scale machine learning. In: *Proc 12th USENIX Conf Oper Syst Des Implement*. 2016.
37. Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. *J Can Dent Assoc*. 2015;**70**(3):156–157. doi:10.1007/s13398-014-0173-7.2
38. Kingma DP, Ba JL. Adam: a method for stochastic gradient descent. In: *ICLR Int Conf Learn Represent*. 2015.
39. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res*. 2014;**15**:1929–1958. doi:10.1214/12-AOS1000
40. Gal Y, *et al*. A Theoretically grounded application of dropout in recurrent neural networks. In: Lee DD, Luxburg UV, eds. *Proceedings of the 30th International Conference on Neural Information Processing Systems*. Red Hook, NY: Curran Associates Inc.; 2016:1027–1035.
41. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)*. 2012;**22**(3):276–282.
42. Phan H, *et al*. SeqSleepNet: end-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging. *IEEE Trans Neural Syst Rehabil Eng*. 2019;**27**(3):400–410.
43. Phan H, et al. Deep Transfer Learning for Single-Channel Automatic Sleep Staging with Channel Mismatch. In: *27th European Signal Processing Conference (EUSIPCO)*. IEEE; 2019:1–5.
44. 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.
45. 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.
46. 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.
47. Van Hout S. The American Academy of Sleep Medicine inter-scorer reliability program: sleep stage scoring Richard S. Rosenberg1. *J Clin Sleep Med*. 2013;**9**(1):81–87. doi:10.5664/jcsm.2350
48. Landis JR, Koch GG. The Measurement of Observer Agreement for Categorical Data. *Biometrics*. 1977;**33**(1):159. doi:10.2307/2529310
49. Jing J, *et al*. Development of expert-level automated detection of epileptiform discharges during electroencephalogram interpretation. *JAMA Neurol*. 2019;**02114**:1–6. doi:10.1001/jamaneurol.2019.3485
50. Ng M, *et al*. Why are seizures rare in rapid eye movement sleep? Review of the frequency of seizures in different sleep stages. *Epilepsy Res Treat*. 2013;**2013**:932790.
51. McKenzie MB, *et al*. Breakthrough spikes in rapid eye movement sleep from the epilepsy monitoring unit are associated with peak seizure frequency. *Sleep*. 2019;**43**. doi:10.1093/sleep/zsz281
52. Manganaro S, *et al*. The need for antiepileptic drug chronotherapy to treat selected childhood epilepsy syndromes and avert the harmful consequences of drug resistance. *J Cent Nerv Syst Dis*. 2017;**9**:1179573516685883.
53. Latreille V, *et al*. Co-morbid sleep disorders and epilepsy: a narrative review and case examples. *Epilepsy Res*. 2018;**145**:185–197.