Deep learning for automated feature discovery and classification of sleep stages

Michael Sokolovsky*, Francisco Guerrero*, Sarun Paisarnsrisomsuk*, Carolina Ruiz*‡, Sergio A. Alvarez†‡
*Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609 USA
†Boston College, 140 Commonwealth Avenue, Chestnut Hill, MA 02467 USA
E-mail for correspondence: alvarez@bc.edu
‡Senior authors

Abstract—Convolutional neural networks (CNN) have demonstrated state-of-the-art classification results in image categorization, but have received comparatively little attention for classification of one-dimensional physiological signals. We design a deep CNN architecture for automated sleep stage classification of human sleep EEG and EOG signals. The CNN proposed in this paper amply outperforms recent work that uses a different CNN architecture over a single-EEG-channel version of the same dataset. We show that the performance gains achieved by our network rely mainly on network depth, and not on the use of several signal channels. Performance of our approach is on par with human expert inter-scorer agreement. By examining the internal activation levels of our CNN, we find that it spontaneously discovers signal features such as sleep spindles and slow waves that figure prominently in sleep stage categorization as performed by human experts.

Index Terms—Clinical neuroscience, sleep apnea, electrophysiology, machine learning, neural networks, feature extraction

1 Introduction

The diagnosis of sleep disorders requires an analysis of physiological information recorded during sleep. Sleep scoring (or staging) is a key ingredient in this analysis, in which continuous-time signals are categorized into sleep stages in discrete time intervals, typically 30s in duration [1]. The technique relies on highly trained human technicians, making the process time-consuming, and yields results that are subject to error, subjectivity, and variation [2], [3]. Automated sleep scoring can, therefore, contribute to more efficient, objective, and reliable diagnosis of sleep-related disorders. Motivated by the excellent performance achieved by convolutional neural networks (CNNs) in the task of image classification [4], [5], [6], we investigate the application of deep CNNs for polysomnographic signal classification in the sleep scoring domain, showing that near-human performance can be attained. We go on to describe some of the characteristic signal features that are learned by such CNNs during training over human sleep signal data.

1.1 Contributions of this paper

This paper makes two main contributions. First, we describe a novel deep CNN network design that is trained to classify multi-channel human polysomnogram signal data into sleep stages, and that improves upon the classification performance of prior CNN work for the same task. We show that this performance gain relies mainly on increased network depth, and not on the use of multi-channel data. Performance of our approach is also competitive with prior work that uses a more complex combination of CNN and LSTM recurrent neural networks (RNN), as well as with human expert inter-scorer performance.

The second, and perhaps more important contribution of this paper, is an initial investigation of the internal representation of the sleep signals that is learned by the proposed CNN. By studying the hidden internal activation levels that occur in response to different types of input signals, both natural and synthetic, we find that the CNN proposed in this paper spontaneously learns specific signal features such as sleep spindles and slow waves that play important roles in sleep staging by human experts. As far as we are aware, this constitutes the first report of such feature emergence in CNN for one-dimensional sleep signal classification.[1]

1.2 Structure of the paper

Section 2 provides general background information about the task of sleep stage classification, and about convolutional neural networks, which will serve as the basis for the classification approach used in the paper. Section 2 also includes a description of related work. Section 3 describes the specific human sleep data set to be used in the experimental evaluation of the proposed approach, as well as the architectural details of the convolutional network model, and the experimental evaluation procedures, including the approach used to identify learned internal features. Section 4 discusses the results obtained, and describes some of the specific signal features that develop within the network during training. The paper ends with conclusions and a description of future work.

1. The present paper is an extended version of [7].

2 Background and Related Work

2.1 Human sleep and sleep stage classification

Sleep is a natural recurring state in humans and many animal species that is known to be vital to human health [8]. Sleep deficiency correlates with health problems, including Alzheimer's disease and dementia [9], [10], disruption of processes associated with inflammatory response and cholesterol regulation [11], Parkinson's disease [12] and heart disease [13], [14].

2.1.1 Polysomnography (PSG)

Polysomnographic sleep studies provide an important tool in diagnosing sleep disorders associated with sleep deficiency. Polysomnography is a multi-parameter test based on several different types of physiological signals. Electrical signals from sensors placed on the body during sleep are recorded and used for analyzing patients' sleep. These signal recordings are collectively called polysomnograms (PSGs). PSGs are composed of data from electroencephalograms (EEG, corresponding to cortical electrical activity), electrooculograms (EOG, corresponding to muscular activity associated with movement of the eyes), electromyograms (EMG, corresponding to muscular activity associated with movement of the chin or legs), electrocardiograms (ECG, corresponding to electrical activity of the heart), as well as other signals corresponding to respiration air flow and blood oxygen levels. The present paper focuses on EEG and EOG signals. Details of the signals utilized appear in Section 3.1.

2.1.2 Standards for human sleep stage scoring

The American Academy of Sleep Medicine (AASM) maintains standards for conducting sleep studies and for categorizing PSG signal data into four sleep stage categories: Sleep Stage 1 (S1), Sleep Stage 2 (S2), Sleep Stage 3 (S3), and Rapid-Eye Movement or REM (R) [15]. Stages 1–3 describe varying depths of sleep. Stages 1 and 2 are often termed light sleep (with Stage 1 being the lighter of the two), while Stage 3 is termed deep sleep. REM sleep owes its name to rapid movements of the closed eyes during sleep, and is associated with dreaming [16], though dreams are now known to occur in other stages of sleep, as well, with evidence that dream content differs between REM and non-REM sleep [17].

During staging, PSG signals are divided into 30-second intervals called sleep epochs, each of which is scored by a technician into either one of the four stages, a wake stage, or a movement stage. The resulting sequences of sleep stages are known as hypnograms. See Fig. 1. Scoring decisions are made by quantitative and visual analysis of electrical signals, relying on spectral characteristics such as low-frequency (delta band) waves in stage S3, as well as time-domain features such as K-complexes in stage S2. The short bursts of periodic waveforms known as sleep spindles that are common in stage S2 represent a third class of mixed time-frequency domain features used during human expert scoring. See Fig. 2. Sleep technicians must be trained to score sleep reliably, and the task of of scoring a sleep study requires considerable effort and time.

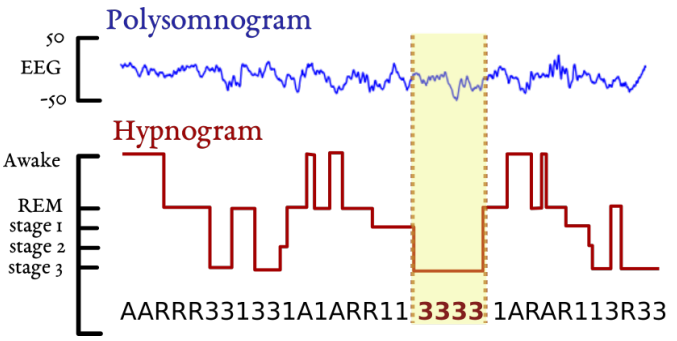


Fig. 1. Sleep scoring involves examining a continuous-time polysomnogram (top) to produce a symbolic, discrete-time hypnogram (bottom). Only one EEG channel is shown. Multiple EEG, EOG and other channels are typically used.

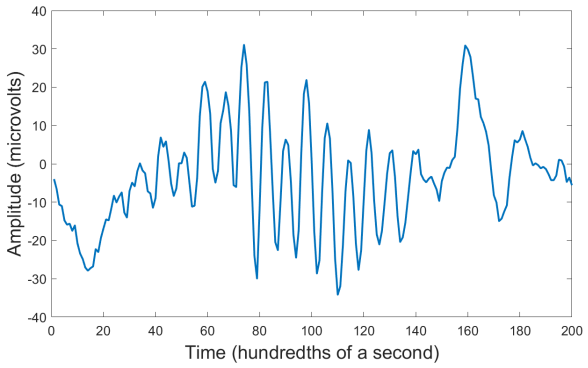


Fig. 2. Sample of stage S2 sleep EEG, showing a sleep spindle. Expert scorers rely on spindles and other time-domain and frequency-domain features to assign sleep samples to stages.

2.2 Convolutional neural networks

Convolutional Neural Networks (CNNs) are variants of neural network (NN) statistical learning models and have been successfully applied to image recognition tasks, achieving current state-of-art results in image classification [18], [6], [4]. Like other neural network types for classification, CNNs take as input unlabeled data in the form of vectors (multivariate measurements), and output information about the predicted class membership of the input vectors.

CNNs are composed of stacks of three main types of processing layers: convolutional layers, pooling layers, and fully connected layers. Each successive layer applies a function to its input data vectors and passes a transformed version of the data vectors as input to the next layer. See Fig. 3. We briefly describe each of these types of layers, below. See [19] for a more detailed discussion of CNNs.

2.2.1 Convolutional layers

Convolutional layers apply linear transformations followed by non-linear activation functions to obtain each individual component of the data vector to be passed to the next layer. The components of a layer's input vector that contribute to a given output component are restricted to a small contiguous patch of the input layer, making the linear transformation a "local" operation. Furthermore, the coefficients of these linear transformations are identical for all output components,

making processing invariant to translations in the input. Because coefficients are shared across processing units, it is common to view the matrix of coefficients as a "filter" (kernel) that slides across the input in order to compute the linear transformation at various locations. Rectified linear (ReLU) activation functions and their variants [20] are typically applied to the outputs of the convolutional filters.

2.2.2 Pooling layers

In contrast with convolutional layers, pooling layers simply down-sample the data vectors from the preceding layer by applying an aggregating function over local patches of contiguous input elements, yielding a lower-resolution rendering of the input vectors. The maximum (over each local patch of the given layer's input) is a typical choice of aggregating function, in what is referred to as "max-pooling". Pooling is intended to control over-fitting by reducing the total number of parameters in the network, and to increase robustness to minor variations in the input data [19]. Pooling also increases filters' receptive fields, allowing filters in deeper layers to use information from a greater portion of each input datum.

2.2.3 Fully connected layers

The final (output) layers of a CNN are most often fully connected layers as in traditional NNs. A fully connected layer applies a linear transformation to its input vectors, followed by a nonlinear activation function. Unlike a convolutional layer, the linear transformation associated with a fully connected layer is not subject to the locality and spatial invariance constraints. In other words, the output of a fully connected layer is the result of applying the nonlinear activation function to a completely general linear transformation of the full input vector. For classification tasks, it is customary to include one output unit per class, and to use a softmax activation function to ensure that the output vector is normalized to sum to 1, reflecting an estimate of the posterior class distribution as suggested in Fig. 3.

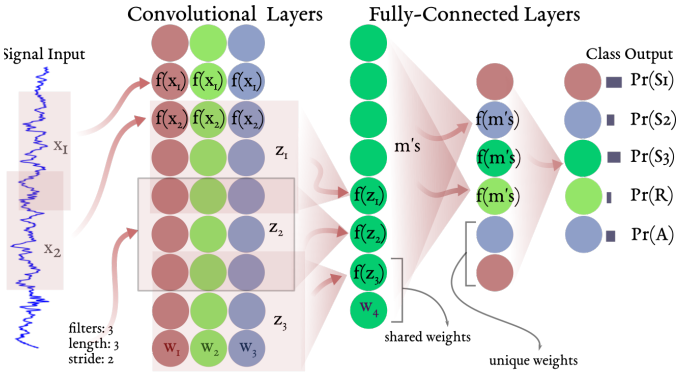


Fig. 3. Convolutional neural network illustration. Pooling layers (see text) not shown.

General comments on CNNs

NNs' predictive power derives from their ability to represent highly nonlinear functions through the composition of nested nonlinear activation functions and linear transformations. While traditional NN layers apply functions to their entire input vectors, the convolutional and pooling layers in CNNs apply functions to smaller local patches of data. The sharing of coefficients across patches, described above, requires fewer parameters and therefore decreases the representational power of CNNs compared to fully-connected NNs of the same number of processing units. To an extent, this provides a useful sort of regularization that helps to reduce overfitting, given the large size of typical deep networks. It also reduces the computation time needed for training models in domains with high-dimensional data, and works well for problems where local patterns are meaningful for classification. The latter situation occurs in sleep scoring, where translation-invariant signal features are important, including K-complexes and sleep spindles [15].

Prior work in visualizing CNN layer outputs in the context of image recognition suggest that CNNs extract local patterns within a data sample in early layers and aggregate them into larger patterns descriptive of the entire sample in deeper layers of the model [21]. This phenomenon of hierarchical feature development with depth has not previously been documented in the context of sleep stage classification based on PSG signals, but will be described in an upcoming paper (see [22] for a preliminary version).

2.3 CNNs, signal classification, and sleep

CNNs are often used for image classification and two-dimensional pattern detection tasks. Less research has been done in one-dimensional signal classification with deep networks. Prior research exists in sound recognition [23], [24], [25] and EEG signal analysis [26]. Deep NN research in sleep stage classification from PSG signals is rarer still. We are aware of prior PSG-related works using auto-encoders [27], deep belief networks [28], and CNN [29], respectively. The recent paper [30] also considers a more complex architecture that combines CNN and LSTM recurrent neural networks. As compared with [29], the present paper explores the use of multi-channel PSG input data in preference to only single-channel EEG, and uses deeper CNN architectures with the aim of improving sleep stage classification performance. Notably, we also take steps toward understanding the source of the superior performance of our network, showing that it spontaneously discovers signal features such as the rapid eye movements that characterize the REM sleep stage, the sleep spindles that occur in sleep stage S2, and the slow waves that are typical of stage S3.

3 Methodology

3.1 Data

PSGs from the publicly-available Physionet database [31] were used for training and evaluating neural network models. Specifically, the Study 1 data from the Sleep-EDF Database [Expanded] [32] was used. The database is composed of 20 healthy patients' PSG data from two full days of recording, totaling 39 recordings (data from one patient was only available for one day). PSG data for each patient consisted of two EEG signals, (EEG Fpz-Cz, EEG Pz-Oz) and one EOG signal (EOG horizontal) sampled at 100 Hz.

Accompanying hypnograms for the full day PSG recordings are included. Each patient record (all samples for one night of that patient's sleep) was labeled by precisely one of six human expert scorers. Each expert scored one or more patient records.

Pre-processing. Input data representation

To avoid including epochs from non-sleep periods, raw data before the first observed epoch of sleep in each PSG recording were discarded, as were the data after the last observed epoch of sleep. Because the stage classification of a given sleep epoch depends on information about neighboring epochs [15], the multi-hour PSG signals were segmented into overlapping 150-second, or 5-epoch, samples as in [29]. The mean number of samples per patient was 1813, with a standard deviation of 357. Models were trained to categorize the middle epoch of each sample. The movement epochs during sleep were removed from the dataset because of their rarity, leaving five options of sleep encoding for each sleep epoch. Given that three signal channels were used, two EEG channels and one EOG channel, input data to the network took the form of two-dimensional data of shape (15000, 3) composed of three 150-second signals sampled at 100 Hz. The first convolutional layer in the model interpreted the three signals as channels of a 15,000-length vector; filters in that layer take act on all three channels at once. The class labels were coded as one-of-five categories (Awake, S1, S2, S3, REM) in a one-hot encoded vector.

3.2 Model architecture

The model architectures explored in this study were inspired by recent state-of-the-art networks that achieve high performance by increasing depth. Architectures modeled on the VGG network [6] and its successors, that use small convolutional filters, multiple stages with stacked convolutional layers separated by pooling layers, and increasing numbers of filters with depth, were tested alongside variants of residual networks [33] that feature skip connections. We experimented with architectures divided into $1-6$ stages, each consisting of $1-6$ convolutional layers, plus batch normalization (which was found to be more effective than dropout and both $L1$ and $L2$ regularization) and max-pooling layers. Several such architectures were considered. We did not perform an exhaustive parameter sweep, as the time needed would be prohibitive. Instead, we adjusted values incrementally, taking into account available GPU memory, among other factors. The model reported in this paper is the current best performing model and is based on the architecture suggestions outlined in [34]. In particular, the architecture is designed so that the receptive fields of the filters in the deepest layers of the network include most of the five-epoch input window. The model structure is shown in Fig. 4 and is described below.

Following [34], our model is structured in levels containing stacked convolutional layers separated by pooling layers. The model contains seven levels with 6 convolutional layers in the first level, 3 convolutional layers in each of the three levels that follow, a single convolutional layer in the fifth level and sixth levels, followed by a final sixth level consisting of two fully-connected layers. All early

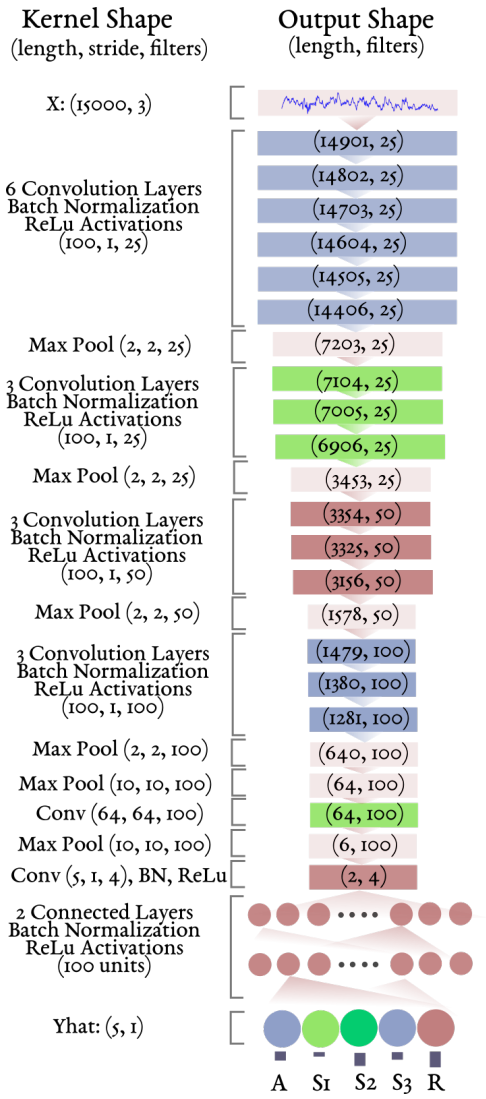


Fig. 4. Network architecture proposed in this paper.

convolutional layers have kernel sizes of length 100 as in [29] and a stride size of 1. Because the sleep epoch to be classified was in the middle of each 150-second input vector, no padding was used during training, and experiments with padding did not yield better results. As in [6], the number of filters in each convolution stayed the same or increased after down-sampling so that each layer within every stage required roughly the same computational time. Activations in all layers were ReLu linear rectifiers, known to provide improved training convergence [18]. The output of the model was a vector of five numbers representing the probability of each class, calculated via a final softmax layer.

3.3 Model training and evaluation

3.3.1 Training procedure

Model weights were initialized as in [33], by sampling from a Gaussian distribution with zero mean and standard deviation $\sqrt{2/n_l}$ in layer l, where n_l is the product of the number of input channels and the number of weights per filter in layer l. Stochastic gradient descent was used to minimize the cross entropy loss function applied to the

softmaxed output. Training used mini batches of size 280 – the largest size that could be accommodated by the P100 GPU. To account for class imbalance, gradient updates from mini batch samples were weighted by the inverse of class frequency in the training set to achieve balanced training across each class. The only regularization used was the addition of batch normalization layers preceding non-linear activations as in [35]. The initial learning rate of 0.01 was progressively decreased after validation accuracy stopped increasing. Models were trained for between 30 and 100 epochs. For ten-fold experiments, training sets consisted of approximately 27,000 samples. Different model architectures were built and trained in TensorFlow [36] and Keras [37] running on the NVIDIA CUDA platform, using NVIDIA K20, K80, and P100 GPUs.

3.3.2 Hyperparameter selection and performance evaluation

An exhaustive search of the hyperparameter space was not performed because of time constraints. Guidance from the literature was used in setting certain hyperparameter values in advance of training. For example, the depth of the CNN model was determined based on receptive field considerations as in the reference by X. Cao [34]. Limited alternate values of other hyperparameters were evaluated by means of four-fold cross-validation. The 20 patients' data were compiled into four folds, each containing training, validation, and test sets; training sets included 13 patients, validation sets included 2 patients, and test sets included 5 patients. Folds were randomly compiled so that every patient's data appeared exactly once in a test set for one of the four folds. Performance on the validation fold was used for hyperparameter selection. For each fold, trained models were evaluated on test data. Final performance metric values were calculated directly from the cumulative confusion matrix created by adding together the confusion matrices from each fold. Metrics reported include precision, recall, and F1-score on each of the five classes as well as net classification accuracy.

3.3.3 Comparison with human expert inter-scorer agreement

Agreement among human expert sleep scorers was also used as a benchmark for classification performance, relying on the inter-scorer data reported in [2], [1], [3].

3.4 Visualization and analysis of learned internal representation

Features learned by the network were studied through a visualization technique based on [38] that displays the difference in the activation of a given filter that results from removing segments of the input signal(s) at different points in time. We modified Algorithm 1 in [38], which is designed for two-dimensional images, to allow its use on the one-dimensional PSG signals considered in the present paper. Human sleep PSG samples were used as input signals to drive the visualization process. Synthetic limited-duration sinusoidal burst signals were also used as inputs to better identify the time-frequency response of particular filters, providing supplementary information to test hypotheses about the nature of the learned features in terms of the amplitude, frequency, and duration of the input signals.

We previously attempted to use a visualization technique based on gradient descent synthesis of signal inputs that lead to maximum activation of a given filter [39]. However, we found it necessary to employ regularization in order to prevent high-frequency signal content that appears unnatural, and were not able to arrive at an objective basis for selecting the value of the regularization constant. An alternate visualization route that merits exploration in future work is the so-called multifaceted visualization of [40]. The latter technique allows uncovering different feature types that may be associated with filters in deeper layers.

4 Results and Discussion

4.1 Predictive performance

4.1.1 Overall and per-stage classification results

TABLE 1
Cumulative confusion matrix with precision, recall, and F1-Score. Table entries show the number of epochs of a given class as classified by a technician (rows) and by the model (columns). S1, S2, S3, A, R are the class labels for sleep stage 1, sleep stage 2, sleep stage 3, Awake, and REM sleep, respectively. Overall accuracy (and micro-F1) is 81%. Macro-F1 is 72% (arithmetic mean of F1 scores in rightmost column).

	S1	S2	S3	A	R	Recall	F1
S1	1402	554	11	356	369	48%	47%
S2	697	15243	777	533	487	87%	87%
S3	9	587	5071	34	1	88%	87%
A	550	142	22	1589	131	56%	56%
R	414	700	12	302	6276	80%	83%
Precision	46%	88%	86%	56%	86%		72%

Overall accuracy for the model was 81%, as calculated from the cumulative confusion matrix in Table 1, by dividing the sum of the diagonal elements (samples classified correctly) by the sum of all of the elements (all samples). Very good individual class performance was attained on sleep stages S2, S3, and REM in terms of precision, recall, and F1-scores. The worst performing classes by the same metrics were S1 and the Awake stage, as reflected, also, in a lowered overall macro-F1 score of 72%. The observed per-stage performance differences are not surprising, as stages S1 and Awake occur much less frequently in normal human sleep than other stages, and stage S2 is the most frequent of all; see [41], p. 20 (keeping in mind that stage S3 in the updated AASM standard used in the present paper corresponds, approximately, to the union of stages 3 and 4 in the older Rechtschaffen-Kales standard used in [41]). Additionally, Awake stage performance was limited by the elimination during pre-processing of all input data occurring before the onset of sleep and all data after the last epoch of sleep, as described in the second paragraph of Section 3.1.

4.1.2 Comparison with prior work

Performance of our model surpasses that of prior CNN work [29] on the same dataset. The latter paper reported an overall accuracy of 74%, training a shallower, six-layer CNN model on a single EEG Fpz-Cz signal. We modified our model to train on a single EEG signal in order to allow a performance comparison with [29]. After evaluating

performance on each fold, total accuracy of our approach on a single EEG channel was 80%, which comfortably exceeds the 74% of [29]. Performance of our approach decreases only minimally when the number of channels is reduced, from 81% for three channels to 80% for a single channel. This fact, together with the substantial performance improvement of our approach over [29], suggests that increased network depth has a greater effect on performance than the use of multiple input channels.

A more complex architecture that uses a CNN as a feature extractor, followed by a bi-directional LSTM recurrent neural network [30] attains a maximum accuracy of 82% on the same data set used in the present paper, only a 1% improvement over our results. In comparing [30] with the present paper, it is important to note that data preprocessing in [30] retains 30-minute periods of Awake epochs both before the onset of sleep and after the end of sleep (see Section IV.A of [30]), while the present paper keeps only the data between sleep onset and the last epoch of sleep. Thus, the CNN model of [30] is provided with a substantially greater amount of Awake data from which to model that particular stage. Better performance of [30] on stage Awake leads to higher macro-F1 scores of 76.9% or 73.1% than our 72% in Table 1, depending on which of the two available EEG channels is used. The present paper goes considerably beyond [30] in uncovering the internal network features that result from learning, and that underlie the network's classification performance. See Section 4.2, where it is also shown that the emergent internal features relate closely to signal features used by human sleep experts.

4.1.3 Comparison with human expert scorers

Comparing our results to human inter-scorer agreement provides a contextual backdrop for the classification performance achieved in this and other papers. Whereas one would expect an individual scorer to be self-consistent in their evaluation of sleep stages, a group of scorers may differ in how they score the same epochs. These rates of inter-scorer agreement would then serve as performance goals for a model trained on many scorers' classifications, such as ours.

Table 2 shows results reproduced from [2], a study that compiled more than 2,500 human scorers' classifications on 1,800 epochs for a total over 3,000,000 scoring decisions. Table 2 can be thought of as a confusion matrix, the rows representing ground truth for each sleep epoch (interpreted as majority score attributed to the epoch) and columns as the classifications of the collective pool of scorers. Overall inter-scorer accuracy was 82.6%, which is on par with our results. Comparing with Table 1, human inter-scorer agreement performs better than our deep CNN on stages S1, Awake, and REM, while it matches our network's classification performance on stage S2; and our network's stage S3 performance exceeds inter-scorer agreement. Human scorers' superiority on S1 and Awake stages, especially, leads to a higher human macro-averaged F1 score of 77% in Table 2, as compared with our network's score of 72% in Table 1.

We hypothesize that the observed differences in recall and precision between our model and [2] arise from the large variance in human inter-scorer agreement and the small number of scorers (only seven) who scored our dataset. High variation between scorers is supported by other research in inter-scorer accuracy. Reviews of inter-scorer agreement and meta studies report similar or worse results for overall inter-scorer accuracy and varying performance across classes, with the highest accuracy from stage REM (78-94%) and sleep stage S2 epochs (79-90%) and the lowest from stage S3 (69%) [1], [3]. Surprisingly, intra-scorer agreement appears to be comparable to inter-scorer agreement; one study reported accuracies of the same scorer on data re-scored after a median of 6.5 months to be in the range of 79-87% [3].

The reported inter-scorer reliability of 82.6% [2] serves as an aspirational benchmark for objectively meaningful model accuracy over many epochs and scorers. Our model's accuracy of 81%, and the 82% accuracy of [30] are on par with this benchmark.

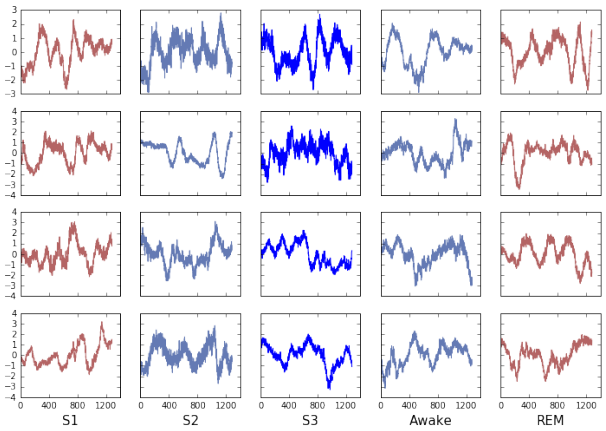


Fig. 5. Sample hidden layer outputs for the five classes. Outputs taken from same filter in final convolutional layer after batch normalization.

TABLE 2
Inter-scorer agreement conducted by the AASM: reproduced from [2] with proxy precision, proxy recall, and accuracy reported. Overall inter-scorer agreement is 82.6%. Macro-averaged F1 is 77%.

	S1	S2	S3	A	R	Recall
S1	187,634	64,700	205	32,545	12,910	63%
S2	116,274	1,570,861	121,141	13,080	23,180	85%
S3	298	87,033	181,337	224	350	67%
A	29,658	11,222	779	250,434	5703	84%
R	25,875	22,883	342	6624	531,611	91%
Prec	52%	89%	60%	83%	93%	
F1	57%	87%	63%	86%	92%	77%

4.2 Learned features

Neural networks are notorious for providing opaque representations of knowledge. Given the very strong predictive performance demonstrated in Section 4.1 by the deep network of the present paper, we are nonetheless motivated to explore the features learned by the hidden layers of the network, which are the basis for its predictions. We approached this task through a combination of techniques, including visualization and analysis of various filters' responses to human sleep as well as to synthetic limited-duration sinusoidal wave bursts. With traditional CNNs, it is common to display layer outputs as two-dimensional

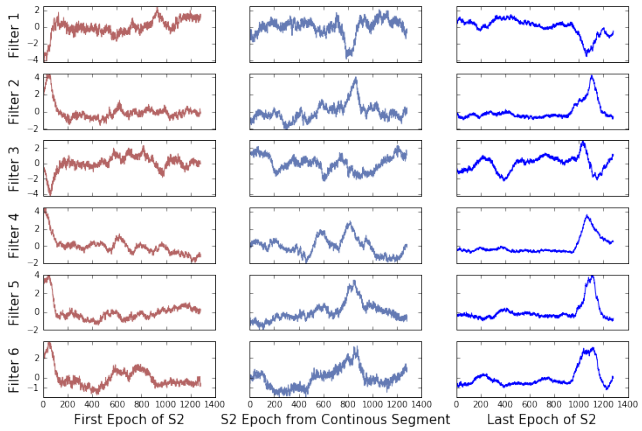


Fig. 6. Hidden layer outputs for three sleep stage S2 datapoints. Outputs taken from random filters in final convolutional layer after batch normalization.

heat maps. Because our data is one-dimensional, we display hidden layer outputs as line graphs.

4.2.1 Visualization of randomly selected signal epochs

Our early visualization attempts used randomly selected input data samples belonging to different sleep stages, in the hope of uncovering differences in hidden layer activations across stages. Unfortunately, most of these visualizations did not reveal any such differences clearly. For example, see Fig. 5, which shows filter activations in the final convolutional layer in the model proposed in the present paper.

The complexity associated with untangling the hidden layer activations stems, in part, from the fact that each input datum comprises not just one sleep epoch, but rather a sequence of five consecutive epochs, with the label of the datum corresponding to the sleep stage of the middle epoch. We denote a specific data point as the composition of the five epochs that comprise it, such as *1333R* or *AA222*, where 1, 2, 3, A, and R are the five class labels. We visualize the hidden layer outputs on data points that share the same class label for the middle epoch but have different labels for the neighboring epochs. Fig. 6 contains hidden layer outputs on three distinct data points all labeled S2: the data point corresponding to the first epoch of sleep stage S2 in a patient, *11222*, a continuous portion from the same patient, *22222*, and the data point corresponding to the last epoch of sleep stage S2, *222AA*. Higher volatility is observed at portions in the signal that correlate positionally with the area of the input data that are not sleep stage S2. These observations suggest that the internal network response differentiates between stages within a five-epoch datum. In combination with the reduced performance on transitional stages reported in the inter-scorer study [2], these visualizations point to categorization differences between continuous and transitional sleep stage epochs as an area for additional exploration. We describe our results in this direction below.

4.2.2 Network discovery of signal features used by human experts

An interesting finding surfaced when we examined how the internal layers of the network respond to localized signal features in human sleep signals, as well as in controlled synthetic signals. Namely, internal features develop spontaneously within the hidden layers of the network during training that are associated closely with specific features that figure prominently in the AASM staging standard, which expert human scorers rely on closely. We describe three of these here briefly. Additional details will be provided in an upcoming paper. See [22] for a preliminary report.

EOG features

Emergent features of interest were found in both the EOG and EEG channels. The first feature, associated with the EOG channel, relates to the rapid eye movements that are the hallmark of the REM sleep stage. Fig. 7 shows the response to stage REM human sleep of a particular filter in layer 14 of the CNN described in this paper. A visualization technique based on [38] is used that displays the difference in activation produced by removing segments of the input signal(s) at different points in time. The activation difference is displayed at the bottom, showing that the segment of high-amplitude EOG waves toward the center of the plot has the greatest impact on the activation of this filter. Activation difference appears to increase slightly before the onset of the high-amplitude EOG, and to persist past the end of this segment; however, this effect is an artifact of the activation difference computation, which averages activations over a short time-window. The response of the filter of Fig. 7 to synthetic data (not shown in the figure) shows greater response to high-duration EOG signals (above 250μV amplitude), which is consistent with the typical EOG signature of rapid eye movements [42].

EEG features: slow-wave sleep

A first emergent EEG feature relates to stage S3 sleep, described in the AASM standard, versions 2.1 and later, as being characterized by "slow" waves with frequencies in the $0.5 - 2$ Hz range and peak-to-peak amplitudes above 75 μV (see the summary of updates in version 2.1 in [43]). Fig. 8 provides an example of the response to a data sample of actual stage S3 human sleep, of a particular filter in layer 13 of the network described in the present paper. The activation difference is displayed at the bottom, showing that the greatest impact in the activation of this filter is produced by the contiguous segment of slow waves toward the center of the plot in the EEG channel shown at the top.

Fig. 9 shows that the frequency response of the filter identified in the preceding paragraph is greatest in the frequency range of approximately $0.56 - 1.8$ Hz, where we include frequencies for which the filter's response is at least one-half of its maximum value. This range matches the range required for slow-wave sleep very closely. Likewise, Fig. 10 shows that the filter's response is greatest for peak-to-peak signal amplitudes above 100 μV. Therefore, both the frequency response and amplitude response of this particular filter closely match the AASM criteria for slow waves in stage S3 sleep.

EEG features: sleep spindles

A second filter, also in layer 13 of the network, is associated with a different emergent EEG feature. This second filter

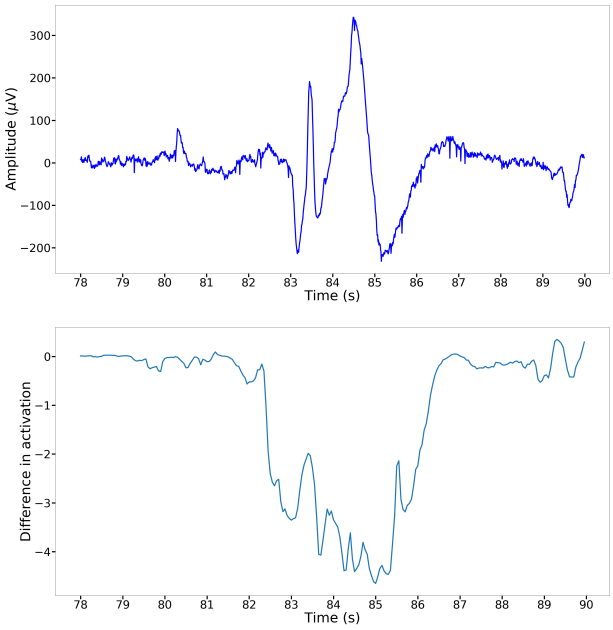


Fig. 7. Response to stage REM human sleep of a filter in layer 14 of the network described in the present paper that learns to detect rapid eye movements. Top plot shows the EOG input channel. Bottom plot shows the difference in filter activation associated with removing a portion of the input at the corresponding locations along the time axis. The difference in activation is greatest for the large-amplitude EOG waves toward the center, associated with eye movements. The displayed activation response starts shortly before the onset of eye movements, and persists past the end of eye movements, because it uses the average over a non-zero time window.

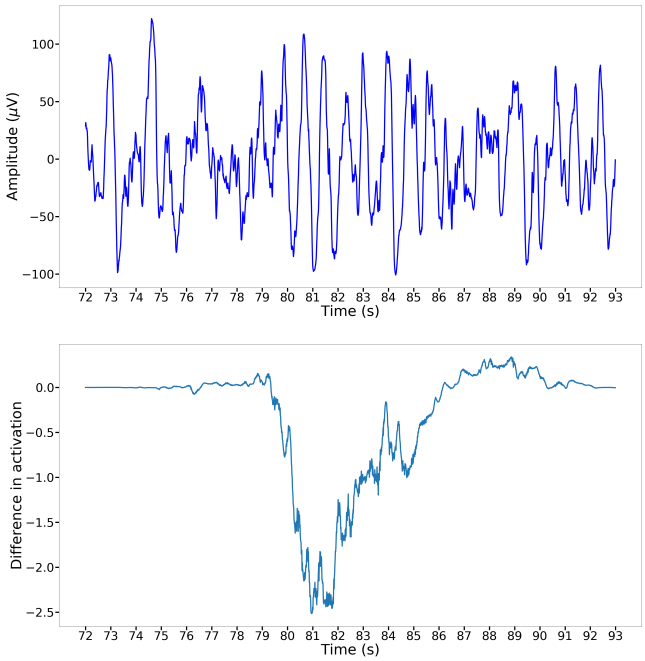


Fig. 8. Response to stage S3 human sleep of a filter in layer 13 of the network described in the present paper that learns to detect slow waves. Top plot shows one EEG input channel. Bottom plot shows the difference in filter activation associated with removing a portion of the input at the corresponding locations along the time axis. The slow wave train toward the center has the greatest effect on activation.

responds most sensitively to the brief $(0.5-2s)$ wave packets at $11-16$ Hz known as sleep spindles, which are characteristic of sleep stage S2, as described in the AASM standard. Fig. 11 shows the response of this filter to a sample of actual stage S2 human sleep that contains a sleep spindle. Fig. 12, Fig. 13, and Fig. 14 show that the activation of this particular filter in response to synthetic limited-duration sinusoidal wave bursts is greatest in the ranges of frequencies, burst durations, and amplitudes that are consistent with sleep spindles.

However, we see in Fig. 12 that, unlike the filter described above that responds almost exclusively to slow-wave sleep, sleep spindles are not the only waveforms detected by the second filter. In fact, this filter also responds to frequency content in the slow-wave sleep range. The apparently different behaviors of the two filters may be attributed to the fact that they both occupy a deep layer. At that point in the network, the desired minimization of the output classification error exerts pressure to model the individual output classes, that is, the sleep stages. Slow waves are strong indicators of stage S3, while sleep spindles alone do not characterize stage S2, as slow waves and K-complexes co-occur with spindles relatively often in stage S2. Thus, it is not surprising that a more complex feature profile should arise in association with sleep spindles, as we observe in the second filter.

4.2.3 *Emergent features: discussion*

We found through an analysis of internal responses of our CNN to human sleep, and confirmed through the use

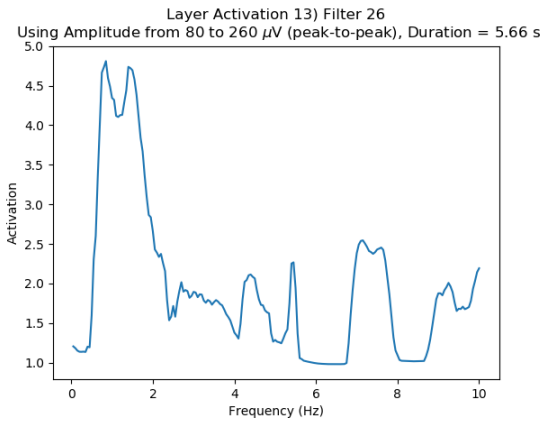


Fig. 9. Frequency response to synthetic data of a filter in layer 13 of the network described in the present paper that learns to detect slow-wave sleep. Activation is greatest in the frequency range $0.56-1.8$ Hz.

of synthetic signal data, the automated emergence during learning of internal features based on the EEG and EOG input signals that correspond closely to those that enter into human expert sleep scoring. A connection of the emergent features with those used by sleep experts is not entirely surprising, as the stage labels provided for supervised training of the network reflect the AASM standard. Nonetheless, it is very interesting that specific filters appear to have specialized in individual features, be it rapid eye movements, slow EEG waves, or sleep spindles, whereas individual features are generally not sufficient to characterize the sleep stages. Given the excellent classification performance of our network in most cases, as discussed in Section 4.1, it

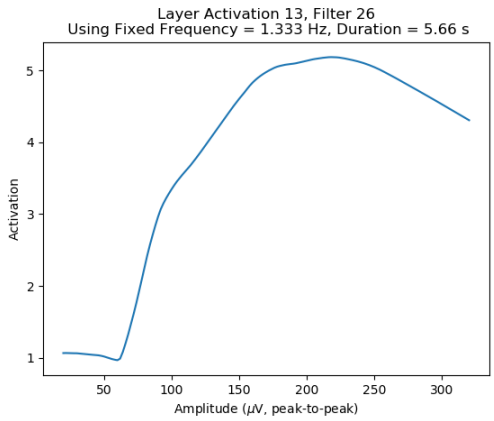


Fig. 10. Signal amplitude response to synthetic data of a filter in layer 13 of the network described in the present paper that learns to detect slow-wave sleep. Activation is greatest for peak-to-peak signal amplitudes above 100 μV.

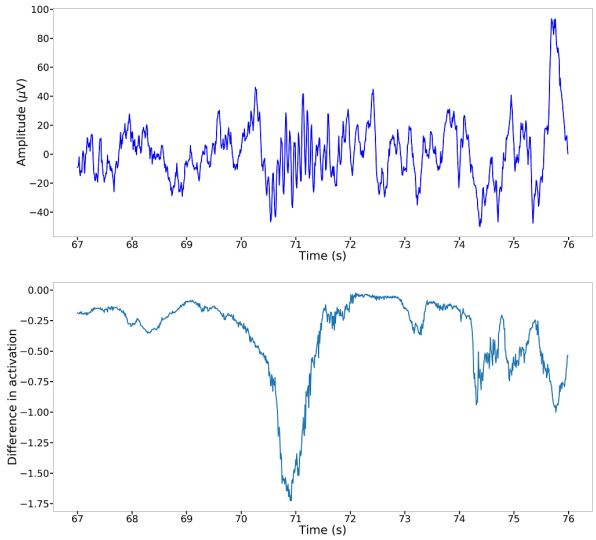


Fig. 11. Response to stage S2 human sleep of a filter in layer 13 of the network described in the present paper that learns to detect sleep spindles. Top plot shows one EEG input channel. Bottom plot shows the difference in filter activation associated with removing a portion of the input at the corresponding locations along the time axis. The spindle toward the center has the greatest effect on activation.

is clear that the learned internal representation captures the characteristics of the corresponding sleep stages well. Therefore, the deepest layers of the network, closest to the output layer, must possess the ability to combine the individual features described above in a suitable manner, reflecting the emergence of more complex features with depth. This phenomenon is reminiscent of the development of a hierarchy of emergent features with depth that has been reported for CNN in two-dimensional image recognition (e.g., [21]). Further results along these lines in the present context of sleep stage classification based on time-dependent physiological signals will be reported in [22].

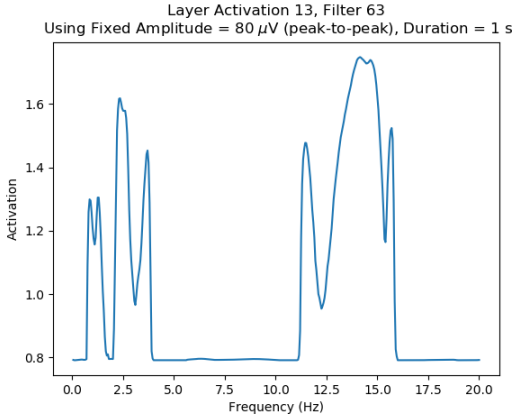


Fig. 12. Frequency response to synthetic data of a filter in layer 13 of the network described in the present paper that learns to detect sleep spindles. Activation is greatest in the spindle frequency range $11 - 16$ Hz, but is also large in the slow wave sleep range near the left end of the plot.

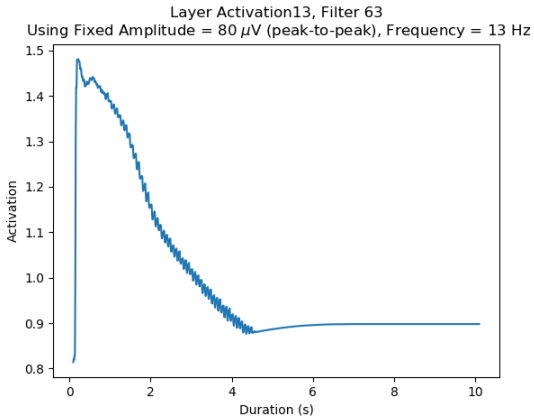


Fig. 13. Duration response to synthetic data of a filter in layer 13 of the network described in the present paper that learns to detect sleep spindles. Activation is greatest in the duration range $0.5 - 2$ s.

5 Conclusions

We explored the use of deep convolutional neural networks (CNN) for classification of sleep stages from multi-channel polysomnogram (PSG) data, more specifically EEG and EOG. Classification accuracy of the proposed technique reached 81%, substantially surpassing that of prior work [29] that uses a single-channel version of the same data set. Reducing the number of input signal channels from three to one reduced accuracy only very slightly, to 80%, thus suggesting that the performance gains of our deeper, multi-channel model over the 74% accuracy reported in [29] derive mainly from increased network depth, and not from the use of multiple signal channels.

Performance of our network design in terms of overall classification accuracy is also competitive with that of a more complex combined CNN and LSTM recurrent neural network approach [30], as well as with human expert inter-scorer agreement [2]. Per-stage performance on stages S1 and Awake is limited by a scarcity of epochs of these stages during normal human sleep, exacerbated by the removal during data preprocessing in the present paper of epochs

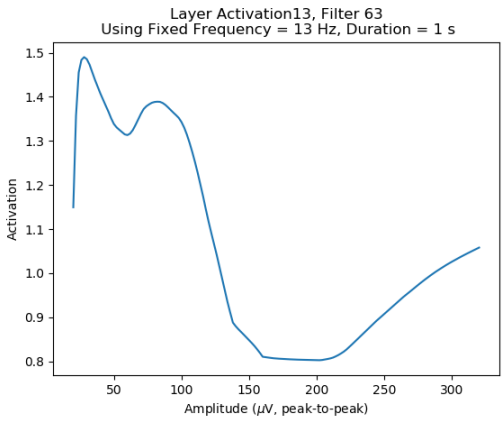


Fig. 14. Signal amplitude response to synthetic data of a filter in layer 13 of the network described in the present paper that learns to detect sleep spindles. Activation is greatest for signal amplitudes in the range $20 - 120$ μV peak-to-peak.

prior to sleep onset and after the end of sleep. This suggests that comparative classification performance of the approach proposed in the present paper may be underestimated by the results reported here.

This paper provides an important contribution beyond a CNN architecture that performs well on the task of automated sleep stage classification. Namely, we also took steps toward understanding the mechanisms through which the proposed deep neural network attains its strong predictive performance. By examining the hidden layers of the network via visualization of activations elicited by both natural and synthetic signals, we showed that individual filters in the network spontaneously learn to identify specific EEG and EOG signal features that figure prominently in human expert scorers' repertoires, such as rapid eye movements, large-amplitude slow waves, and sleep spindles. To the best of our knowledge, feature emergence in deep CNN for one-dimensional sleep signal classification has not been described in prior work. A more detailed description of this phenomenon is the subject of ongoing research by the authors of the present paper. A preliminary report that includes additional results appears in [22]. In addition to better understanding the emergence of well-known signal features in deep networks, the search for predictive learned features that are not currently known to clinicians suggests itself as an interesting direction for future work.

Acknowledgments

The authors thank Majaz Moonis, M.D., of the U. of Massachusetts Medical School, for helpful conversations on polysomnography and sleep.

References

[1] M. H. Silber, S. Ancoli-Israel, M. H. Bonnet, S. Chokroverty, M. M. Grigg-Damberger, M. Hirshkowitz, S. Kapen, S. A. Keenan, M. H. Kryger, T. Penzel *et al.*, "The visual scoring of sleep in adults," *J Clin Sleep Med*, vol. 3, no. 2, pp. 121–131, 2007.

[2] R. S. Rosenberg, S. Van Hout *et al.*, "The American Academy of Sleep Medicine inter-scorer reliability program: sleep stage scoring," *J Clin Sleep Med*, vol. 9, no. 1, pp. 81–87, 2013.

[3] M. M. Grigg-Damberger, "The AASM scoring manual: a critical appraisal," *Current opinion in pulmonary medicine*, vol. 15, no. 6, pp. 540–549, 2009.

[4] W. Rawat and Z. Wang, "Deep convolutional neural networks for image classification: A comprehensive review," *Neural Comput.*, vol. 29, no. 9, pp. 2352–2449, Sep. 2017. [Online]. Available: https://doi.org/10.1162/neco_a_00990

[5] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[6] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[7] M. Sokolovsky, F. Guerrero, S. Paisarnsrisomsuk, C. Ruiz, and S. Alvarez, "Human expert-level automated sleep stage prediction and feature discovery by deep convolutional neural networks," in *Proc. BIOKDD 2018*, Aug. 2018.

[8] M. Kryger, T. Roth, and W. Dement, *Principles and Practice of Sleep Medicine*. Elsevier, 2016.

[9] K. E. Sprecher, R. L. Koscik, C. M. Carlsson, H. Zetterberg, K. Blennow, O. C. Okonkwo, and et al., "Poor sleep is associated with CSF biomarkers of amyloid pathology in cognitively normal adults," *Neurology*, vol. 89, no. 5, pp. 445–453, 2017.

[10] Y.-E. S. Ju, S. J. Ooms, C. Sutphen, S. L. Macauley, M. A. Zangrilli, G. Jerome, and et al., "Slow wave sleep disruption increases cerebrospinal fluid amyloid- levels," *Brain*, vol. 140, pp. 2104–2111, 2017.

[11] V. Aho and et al., "Prolonged sleep restriction induces changes in pathways involved in cholesterol metabolism and inflammatory responses," *Scientific Reports*, vol. 6, no. 24828, 2016, doi:10.1038/srep24828.

[12] K. Wulff, S. Gatti, J. G. Wettstein, and R. G. Foster, "Sleep and circadian rhythm disruption in psychiatric and neurodegenerative disease," *Nature Reviews Neuroscience*, vol. 11, no. 8, pp. 589–599, 2010.

[13] D. J. Gottlieb, S. Redline, F. J. Nieto, C. M. Baldwin, A. B. Newman, H. E. Resnick, and N. M. Punjabi, "Association of usual sleep duration with hypertension: the Sleep Heart Health Study," *Sleep*, vol. 29, no. 8, p. 1009, 2006.

[14] V. K. Somers, D. P. White, R. Amin, W. T. Abraham, F. Costa, A. Culebras, S. Daniels, J. S. Floras, C. E. Hunt, L. J. Olson *et al.*, "Sleep apnea and cardiovascular disease," *Circulation*, vol. 118, no. 10, pp. 1080–1111, 2008.

[15] R. B. Berry, R. Brooks, C. E. Gamaldo, S. M. Harding, C. Marcus, and B. Vaughn, "The AASM manual for the scoring of sleep and associated events. rules, terminology and technical specifications," *American Academy of Sleep Medicine*, 2012.

[16] E. Aserinsky and N. Kleitman, "Regularly occurring periods of eye motility, and concomitant phenomena, during sleep," *Science*, vol. 118, no. 3062, pp. 273–274, 1953.

[17] P. McNamara, P. Johnson, D. McLaren, E. Harris, C. Beauharnais, and S. Auerbach, "REM and NREM sleep mentation," *International Review of Neurobiology*, vol. 92, pp. 69–86, 2010.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[19] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ser. ICCV '15. Washington, DC, USA: IEEE Computer Society, 2015, pp. 1026–1034. [Online]. Available: http://dx.doi.org/10.1109/ICCV.2015.123

[21] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision (ECCV)*. Springer, 2014, pp. 818–833.

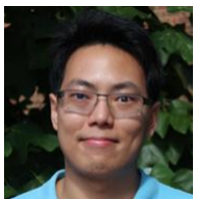
[22] S. Paisarnsrisomsuk, M. Sokolovsky, F. Guerrero, C. Ruiz, and S. Alvarez, "Deep Sleep: convolutional neural networks for predictive modeling of human sleep time-signals," in *KDD 2018 Deep Learning Day*, Aug. 2018.

[23] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *Machine Learning for Signal Processing (MLSP), 2015 IEEE 25th International Workshop on*. IEEE, 2015, pp. 1–6.

[24] S. Dieleman and B. Schrauwen, "End-to-end learning for music audio," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 6964–6968.
[25] H. Zhang, I. McLoughlin, and Y. Song, "Robust sound event recognition using convolutional neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 559–563.
[26] S. Kiranyaz, T. Ince, and M. Gabbouj, "Real-time patient-specific ECG classification by 1-d convolutional neural networks," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 3, pp. 664–675, 2016.
[27] O. Tsinalis, P. M. Matthews, and Y. Guo, "Automatic sleep stage scoring using time-frequency analysis and stacked sparse autoencoders," *Annals of biomedical engineering*, vol. 44, no. 5, pp. 1587–1597, 2016.
[28] M. Längkvist, L. Karlsson, and A. Loutfi, "Sleep stage classification using unsupervised feature learning," *Advances in Artificial Neural Systems*, vol. 2012, 2012.
[29] O. Tsinalis, P. M. Matthews, Y. Guo, and S. Zafeiriou, "Automatic sleep stage scoring with single-channel EEG using convolutional neural networks," *arXiv preprint arXiv:1610.01683*, 2016.
[30] A. Supratak, H. Dong, C. Wu, and Y. Guo, "DeepSleepNet: a model for automatic sleep stage scoring based on raw single-channel EEG," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 11, pp. 1998–2008, 2017.
[31] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "Physiobank, physiotoolkit, and physionet," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.
[32] B. Kemp, A. Zwinderman, B. Tuk, H. Kamphuisen, and J. Oberyé, "Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG," *IEEE-BME*, vol. 47, no. 9, pp. 1185–1194, 2000.
[33] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.
[34] X. Cao, "A practical theory for designing very deep convolutional neural networks." [Online]. Available: https://kaggle2.blob.core.windows.net/forum-message-attachments/69182/2287/A%20practical%20theory%20for%20designing%20very%20deep%20convolutional%20neural%20networks.pdf
[35] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.
[36] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.
[37] F. Chollet, "Keras," 2015.
[38] L. M. Zintgraf, T. S. Cohen, T. Adel, and M. Welling, "Visualizing deep neural network decisions: Prediction difference analysis," *CoRR*, vol. abs/1702.04595, 2017. [Online]. Available: http://arxiv.org/abs/1702.04595
[39] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson, "Understanding neural networks through deep visualization," in *Deep Learning Workshop, International Conference on Machine Learning (ICML)*, 2015.
[40] A. Nguyen, J. Yosinski, and J. Clune, "Multifaceted Feature Visualization: Uncovering the Different Types of Features Learned By Each Neuron in Deep Neural Networks," in *Workshop on Visualization for Deep Learning, International Conference on Machine Learning (ICML)*, Jun. 2016.
[41] M. Carskadon and W. Dement, "Normal human sleep: An overview," in *Principles and Practice of Sleep Medicine*, M. Kryger, T. Roth, and W. Dement, Eds. Elsevier Saunders, 2005, p. 1323.
[42] M. Brown, M. Marmor, Vaegan, E. Zrenner, M. Brigell, and M. Bach, "Iscev standard for clinical electro-oculography (eog) 2006," *Documenta Ophthalmologica. Advances in Ophthalmology*, vol. 113, no. 3, p. 205212, 2006. [Online]. Available: http://doi.org/10.1007/s10633-006-9030-0
[43] R. B. Berry et al., "(updates to) the AASM manual for the scoring of sleep and associated events." [Online]. Available: https://aasm.org/clinical-resources/scoring-manual/

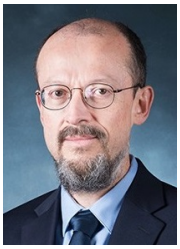
Michael Sokolovsky received an M.S. in Computer Science from Worcester Polytechnic Institute, Worcester, MA, USA, with a thesis on deep neural network models of sleep. He also holds B.A. and B.S. degrees from Brown University, Providence, RI, USA.

Francisco Guerrero received an M.S. in Computer Science from Worcester Polytechnic Institute, Worcester, MA, USA, with a thesis on visualization of neural networks. He previously received an Engineering degree from the U. San Francisco de Quito, Ecuador.



Sarun Paisarnsrisomsuk is a Ph.D. student at Worcester Polytechnic Institute (WPI), Worcester, MA, USA, where he is working on machine learning models of sleep. He received an M.S. in Computer Science from the U. of Virginia, Richmond, VA, USA, and a B.S. in Computer Science from WPI.

Carolina Ruiz received a Ph.D. in Computer Science from the University of Maryland College Park, MD, USA. She is Associate Department Head and Associate Professor in Computer Science at Worcester Polytechnic Institute (WPI), Worcester, MA, USA. She is a Core Faculty Member in Bioinformatics and Computational Biology, Data Science, Neuroscience, and the Healthcare Delivery Institute at WPI, and leads the Knowledge Discovery and Data Mining Research Group there.



Sergio A. Alvarez received a Ph.D. in Applied Mathematics from the University of Maryland College Park, MD, USA, and did postdoctoral work at Carnegie Mellon University. He is currently Chair and Associate Professor in the Department of Computer Science at Boston College, Chestnut Hill, MA, USA, where he heads the Machine Learning Lab. Dr. Alvarez has published over 50 research papers, with a focus on data mining and machine learning, particularly as applied to clinical health.