Deep Sleep: Convolutional Neural Networks for Predictive Modeling of Human Sleep Time-Signals

Sarun Paisarnsrisomsuk
Worcester Polytechnic Institute
Worcester, Massachusetts, USA
spaisarnsrisomsu@wpi.edu

Michael Sokolovsky
Worcester Polytechnic Institute
Worcester, Massachusetts, USA
sokolovsky@wpi.edu

Francisco Guerrero
Worcester Polytechnic Institute
Worcester, Massachusetts, USA
afguerrerohernan@wpi.edu

Carolina Ruiz*
Worcester Polytechnic Institute
Worcester, Massachusetts, USA
ruiz@wpi.edu

Sergio A. Alvarez*†
Boston College
Chestnut Hill, Massachusetts, USA
alvarez@bc.edu

ABSTRACT

We discuss our work on predictive modeling in sleep medicine using deep learning, with attention to domain interpretation of the emergent internal features. More specifically, we describe a high-performing deep convolutional neural network (CNN) architecture for classification of human sleep EEG and EOG signals into sleep stages, the classification performance of which amply exceeds that of recently published CNN work that uses a single EEG channel. We show that the use of multiple signal channels accounts for only a minor fraction of our network's performance, and that the bulk of its performance superiority relies on its greater depth as compared with alternate architectures. By visualizing the response profiles of internal layers of our network to carefully selected input signals after supervised learning, we show that the convolutional filters in these layers extract signal features that closely resemble those used by human sleep experts in manually classifying sleep into stages, and that combine both time-domain and frequency-domain elements. We go on to describe the development of these features with layer depth, showing that the network assembles them in stages, by extracting simple building blocks in shallow layers and combining them in deeper layers to form more complex features. This phenomenon of hierarchical feature formation is well-known in two-dimensional image classification using deep networks, but has not previously been reported in sleep stage classification based on time-varying one-dimensional signals.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; *Supervised learning by classification*; • **Applied computing** → **Health informatics**;

*Senior authors.
†Corresponding author.

KDD'18 Deep Learning Day, August 2018, London, UK

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Undiagnosed sleep apnea is estimated to have an economic impact on the order of $150 billion per year in the US alone,[1] and sleep disorders are associated with a variety of serious health problems [Medic et al. 2017]. The diagnosis of sleep disorders relies on the process of sleep staging (also known as sleep scoring) [Silber et al. 2007], in which highly trained human experts classify 30-second segments of continuous polysomnography (PSG) signals measured by sensors attached to the body during sleep to a symbolic, discrete-time sequence of sleep stages known as a hypnogram (Fig. 1). Human experts base their scoring decisions on spectral and time-domain features of the PSG signals, including slow waves, sleep spindles, and K-complexes. See Fig. 2.

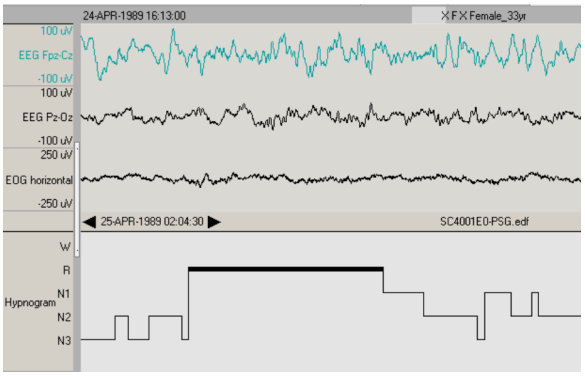

Figure 1: Sleep scoring maps continuous-time polysomnogram data (top) to a symbolic, discrete-time hypnogram. Screen shot of human sleep data from this paper, viewed in Polyman [Roessen and Kemp 2018].

Automated sleep stage classification can contribute to more efficient and reliable diagnosis of sleep-related disorders. In particular, convolutional neural networks (CNN) have recently been applied very successfully to the task of automated sleep stage classification [Sokolovsky et al. 2018; Supratak et al. 2017; Tsinalis et al. 2016b]. Sleep stage classification performance of CNN is now comparable to that of human experts [Sokolovsky et al. 2018]. CNNs'

[1] https://aasm.org/advocacy/initiatives/economic-impact-obstructive-sleep-apnea/

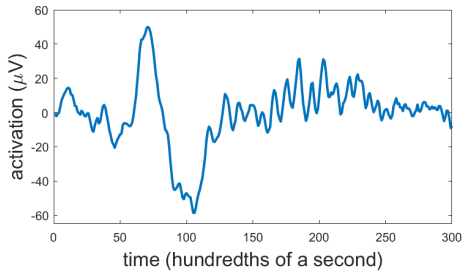

Figure 2: Sample of stage S2 sleep EEG, showing a K-complex followed by a sleep spindle. Total duration is 3s. Expert scorers rely on these and other time-domain and frequency-domain features to assign sleep samples to stages.

impressive predictive performance is, unfortunately, accompanied by opaqueness of their internal mechanisms.

Contributions of this paper

We describe a deep CNN network design that is trained to classify multi-channel human polysomnogram signal data into sleep stages. The performance of our approach compares favorably with that of prior work, as well as with human expert inter-scorer agreement. We find among the internal features that emerge in our deep CNN several that play key roles in the staging process used by human experts, namely sleep spindles and K-complexes, large-amplitude slow waves, and smaller-amplitude alpha waves. These features develop in the shallower and intermediate-depth layers of the CNN and are refined and combined in deeper layers to form more complex features that describe individual sleep stages. Portions of this work will appear in [Sokolovsky et al. 2018]. This paper provides the first demonstration of which we are aware of the emergence of a hierarchy of complex features with increasing depth for sleep EEG signals. Our results may pave the way for the identification of predictive physiological signal features not previously considered by sleep experts. Such features could lead to improvements in the description of sleep progression and the diagnosis of sleep disorders.

2 BACKGROUND

2.1 Polysomnography and sleep stage scoring

Polysomnography (PSG) comprises several types of continuous physiological time signals captured by various sensors attached to the body during all-night sleep studies, including electroencephalography (EEG), which uses electrodes placed on the scalp to measure electrical signals originating in the brain; electrooculography (EOG), which measures movements of the eyes; electrocardiography (ECG), which measures electrical activity of the heart; and electromyography (EMG), which measures muscle contractions. Respiration and blood oxygen levels are typically monitored, as well. These signals are sampled, usually at 100 Hz or higher, resulting in multi-channel time series data. Traditionally, highly trained human experts convert the sampled PSG signals visually to a symbolic representation in terms of a small number of sleep stages, via a classification process known as sleep staging or sleep scoring [Silber et al. 2007].

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, following guidelines such as those provided by the American Association for Sleep Medicine (AASM) [Berry et al. 2012]. Standard stages include the lighter sleep stages S1 and S2, deep sleep stage S3, and the Rapid Eye movement (REM) stage. Wakefulness is often considered to be a stage of sleep, as well, since episodes of wakefulness during sleep are not uncommon. The sequence of sleep stages that results from a PSG recording is known as a hypnogram. See Fig. 1. Scoring decisions are made by quantitative and visual analysis of PSG 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.

2.2 Deep convolutional neural networks for sleep stage classification

Sleep technicians must be trained to score sleep reliably, and the task of of scoring a sleep study requires considerable effort and time. Different scoring experts do not always agree [Grigg-Damberger 2009; Silber et al. 2007]; average inter-scorer agreement has been reported as 82.6% [Rosenberg et al. 2013]. Automated sleep scoring can contribute to more efficient and reliable diagnosis of sleep-related disorders than is possible by human experts. Automated sleep scoring is usually cast as a supervised learning task in which the input consists of epochs (contiguous segments, typically 30s in duration) of the near-continuous sampled PSG signals, and the target is the corresponding sleep stage label of the input epoch.

Several types of deep neural networks have been considered for analysis of PSG sleep signals, including auto-encoders [Tsinalis et al. 2016a], deep belief networks [Längkvist et al. 2012], and CNN [Sokolovsky et al. 2018; Supratak et al. 2017; Tsinalis et al. 2016b]. Classification performance of the latter approaches is now comparable to that of human experts [Sokolovsky et al. 2018]. Deep neural networks have also been applied successfully to other one-dimensional signal classification tasks, including sound recognition [Dieleman and Schrauwen 2014; Piczak 2015; Zhang et al. 2015] and ECG signal analysis [Kiranyaz et al. 2016]. CNNs are known to provide excellent performance in image recognition tasks such as image classification [Krizhevsky et al. 2012; Rawat and Wang 2017; Ren et al. 2015; Simonyan and Zisserman 2014].

CNNs' predictive performance rests on their representation of highly nonlinear functions as the composition of nested nonlinear activation functions and linear transformations that are insensitive to translations of the input in space or time [Goodfellow et al. 2016]. Translation-invariant features are important in image and signal processing tasks, as significant motifs or characteristics may occur in various locations within an input datum. CNN include several types of processing layers, each of which applies a transformation to the preceding layer's vector of activation levels in order to produce a new vector that becomes the input to the next layer. Three main types of transformations are used: linear convolutions that

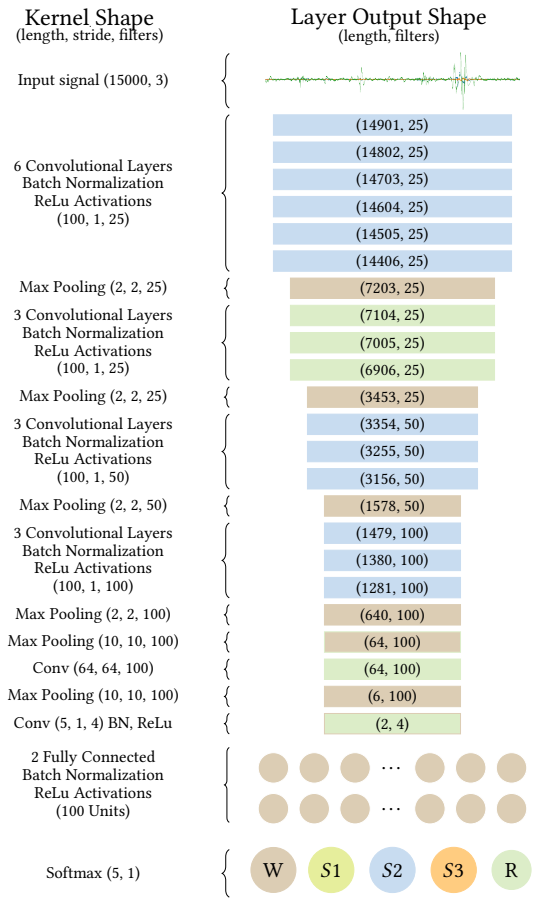

Figure 3: Deep CNN architecture in the present paper.

compute translation-invariant operations that act locally on portions of the preceding layer's activation vector; nonlinear activation transformations that are often applied after the convolution layers; and nonlinear pooling layers that aggregate local regions in the preceding layer's activation vector to achieve greater insensitivity to perturbations of the input. Suitable augmentation of the input data can provide further invariance [Kauderer-Abrams 2018].

3 METHODOLOGY

3.1 CNN classification model

Architectures modeled on the VGG network [Simonyan and Zisserman 2014] 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 [He et al. 2015] that feature skip connections. The model reported in this paper (Fig. 3) is the current best performing model. The architecture is designed, as in [Cao [n. d.]], so that the receptive fields of the filters in the deepest layers include most or all of the five-epoch input window. It consists of a stack of 17 convolutional layers separated by pooling layers, topped by two fully-connected layers that serve as a classifier based on the features extracted by the convolutional structure. All early convolutional layers have kernel sizes of length 100 and a stride size of 1. Because the sleep epoch to be classified is 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 [Simonyan and Zisserman 2014], the number of filters in each convolutional layer increases after down-sampling so that each layer required roughly the same computational time. Activations in all layers were ReLu linear rectifiers, known to provide improved training convergence [Krizhevsky et al. 2012]. The output of the model was a vector of five numbers representing the probability of each class, calculated via a final softmax layer.

3.2 Model training and evaluation

3.2.1 Training procedure. Model weights were initialized as in [He et al. 2015], 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. Training used mini batches of size 280. To account for class imbalance, gradient updates from mini-batch samples were weighted by the inverse of class frequency in the training set. The only regularization used was the addition of batch normalization layers preceding non-linear activations as in [Ioffe and Szegedy 2015]. 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. CNN architectures were built and trained in TensorFlow [Abadi et al. 2016] and Keras [Chollet et al. 2015] on the NVIDIA CUDA platform, using NVIDIA K20, K80, and P100 GPUs.

3.2.2 Cross-validation. Models were first trained using four-fold cross validation, and the best-performing model was retrained using ten-fold cross validation. For the first tier of training, 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. The best-performing model was retrained using ten-fold cross validation. Similar to the first training procedure, folds were randomly compiled so that every patient's data appeared exactly once in a test set for one of the ten folds. Training sets contained 15 patients, validation sets contained 3 patients, and test sets contained 2 patients.

For each fold, trained models were evaluated on test data. Final performance metrics 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.

3.2.3 Bootstrap confidence intervals. Confidence intervals of two standard deviations were calculated and reported on each metric using bootstrap sampling as in [Tsinalis et al. 2016b]. Intervals were calculated from 1000 bootstrapped samples created from the patients' data by the method described below. For example, one of the 20 patients, patient x, was selected at random, and a bootstrapped sample was created from their data. From the final ten

trained models, the bootstrapped sample was fed to the one model in which patient x was in the test set, and a confusion matrix were generated from the output. This process was repeated 1000 times to generate 1000 bootstrapped confusion matrices. Metric averages were calculated from the cumulative confusion matrix consisting of the sum of the 1000 bootstrapped confusion matrices. Lower and upper bounds with a radius of two standard deviations were calculated around each metric's sample mean.

3.2.4 Comparison with human expert inter-scorer agreement. Agreement among human expert sleep scorers was used as a classification benchmark, relying on inter-scorer data from [Grigg-Damberger 2009; Rosenberg et al. 2013; Silber et al. 2007].

3.3 Data

3.3.1 Human sleep data. The publicly-available Physionet database [Goldberger et al. 2000] was used as a source of PSG records for training and evaluating the CNN model described in section 3.1. Specifically, we used the Study 1 data from the Sleep-EDF Database [Expanded] [Kemp et al. 2000]. The database contains PSG recordings for 20 patients, over two full days of recording, totaling 39 single-day data files (data from one patient was only available for one day). PSG data for each patient consists of two EEG signals with electrode placements EEG Fpz-Cz, EEG Pz-Oz, and one EOG signal (EOG horizontal) sampled at 100Hz. Accompanying hypnograms (class labels) for the full day PSG recordings are included, with each day of recording scored by one of six human expert scorers. These are 24-hour recordings that include much non-sleep data. Raw signal data corresponding to wakefulness prior to sleep onset were removed, as were data corresponding to wakefulness after the last scored sleep epoch. This has the advantage of focusing on sleep periods, but the disadvantage of making available only a very small amount of waking data for modeling purposes.

The PSG input signals were segmented into 150-second (5-epoch) samples, in order to include information from neighboring epochs. The scored stage label of the middle epoch of each sample was used as the classification target. Movement epochs were removed from the dataset because of their rarity, leaving five possible stage labels for each epoch: Wake (W), S1, S2, S3, REM. Wake epochs prior to the onset of sleep were also removed. Since 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 100Hz. The first convolutional layer in the model interpreted the three signals as channels of a 15,000-length vector; filters in that layer act across all three channels. Class labels were W, S1, S2, S3, REM.

3.3.2 Synthetic data. Limited-duration, fixed frequency sinusoidal bursts of various amplitudes were employed to better capture the network response to a variety of burst signals that occur in sleep EEG, including $0.5-2$s sleep spindles in stage S2 and $2-3$s alpha bursts in stage REM [Cantero and Atienza 2000]. Delta waves in stage S3 are also of finite duration and can be modeled as bursts. While the K-complexes that occur in stage S2 have a somewhat different time profile, their overall spectral contribution to sleep EEG is associated with slow delta oscillations [Amzica and Steriade 1997] that can be approximated by delta-band bursts.

Time duration of synthetic bursts was limited by means of a smooth envelope, providing slightly better frequency localization as compared with hard time-limiting using a discrete window. Preliminary tests showed that the spectral artifacts that occur with hard time-limiting become particularly problematic for short time durations, introducing spurious responses in the slow-wave frequency range. This fact makes hard-limiting a poor choice for exploring network response to slow-wave bursts. Instead, we selected a soft-limited (smoothed) time window approach. The soft-limited synthetic burst corresponding to frequency f, duration d, and amplitude a is as in Eq. 1. The factor 1.08 is a normalization constant that yields approximately the same L^2 energy as a hard-limited burst with the same parameter values.

$$s_{f,d,a}(t) = 1.08a \sin(2\pi f t) e^{-\left(\frac{2t}{d}\right)^6} \tag{1}$$

The wave form in Eq. 1 is similar to a Gabor-Morlet wavelet (Eq. (1.27) in [Gabor 1946]), but with a more rapidly decaying envelope. Unmodified Gabor-Morlet wavelets have poor temporal localization in practice (difficulty assessing the effective duration of a burst), which in the present context outweighs their superior frequency localization. A comparison of the time and spectral density plots of the smoothed burst signals of Eq. 1 and hard-limited (truncated) bursts appears in Fig. 4. As a rule of thumb, we found that frequency-response results obtained using the burst in Eq. 1 are reliable for burst durations, d, that match or exceed the oscillatory period, $1/f$. This "admissibility condition" may be expressed as follows:

$$fd \geq 1 \tag{2}$$

Eq. 2 coincides with Gabor's time-frequency version of the Heisenberg uncertainty principle (Eq. (1.26) in [Gabor 1946]) that asserts that the product of the uncertainties in time and in frequency is bounded below by $1/2$, if the uncertainties in time and frequency are taken as $d/\sqrt{2}$ and $f/\sqrt{2}$, respectively. As an example, Eq. 2 indicates that attention should be focused on frequencies above 1 Hz when interpreting responses to 1-second bursts; responses to 1s bursts of frequencies below 1 Hz should be interpreted cautiously. This is natural: the latter bursts contain less than one full cycle.

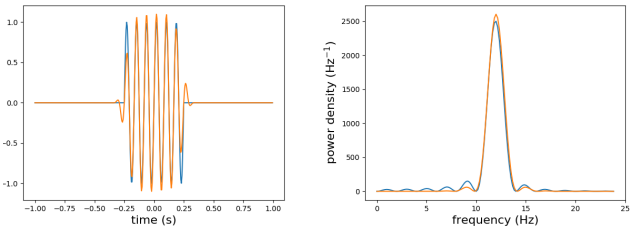

Figure 4: Hard- (blue) and soft-limited (orange) sinusoidal bursts. Time plot (left) shows good time-localization. Power spectral density plot (right) shows reduced spectral leakage (side lobes) for soft-limited bursts, which is desirable.

3.4 Identification of emergent features

Features learned by the network were identified and explored through a combination of visualization and analysis approaches. Human sleep samples were used as input signals to initiate the process, by determining each filter's response to the different sleep stages. See Algorithm 1. Recall that each input datum consists of five epochs as in section 3.3.1, and that a filter's response $\phi(s)$ on network input s is a time-series $\phi(s)(t)$. Filters were associated with a particular sleep stage if 75% or more of their top-activating input signals (elements of S^* in Algorithm 1) belong to that stage.

Algorithm 1: Stage distribution computation

Function *compute stage distribution(filter of interest, ϕ)*
  forall *human sleep data samples, s, in data set* **do**
    use sample s as input to network;
    $\phi(s)$ = activation of filter ϕ (this is a time-series);
    $\phi'(s) = \max_t \phi(s)(t)$ for t in middle epoch;
  end
  C = 90-th percentile of values $\phi'(s)$ for all samples, s;
  S^* = the set of s for which $\phi'(s) \geq C$;
  return sleep stage distribution of signals in S^*;

A technique for visualizing differences in activation based on [Zintgraf et al. 2017] was used to gauge the impact of particular input segments, thus aiding in the identification of input features associated with a given filter. This approach displays the difference in the activation of a filter that results from removing segments of the input signal(s) at different points in time. We modified Algorithm 1 of [Zintgraf et al. 2017], which is designed for two-dimensional images, to allow its use on one-dimensional PSG signals as in the present paper. The basic idea is to measure the difference in activation that results from replacing a portion of the input signal with a random sample from neighboring points. See Algorithm 2, below, in which $N(\mu, \sigma^2)$ denotes a Gaussian probability distribution with the given mean and variance; the range of t reflects the fact that each input signal includes 150s of data, sampled at 100 Hz.

Algorithm 2: Computation of activation differences

Function *activation difference(input signal i, filter of interest ϕ, window width w, sampling width n)*
  forall t *in* $[0, 15000]$ *(steps of* 5*)* **do**
    $i' = i$;
    portion to be removed = $i[t - \frac{w}{2}, t + \frac{w}{2}]$;
    neigh = $i[t - \frac{w}{2} - \frac{n}{2}, t - \frac{w}{2}) \cup i(t + \frac{w}{2}, t + \frac{w}{2} + \frac{n}{2}]$;
    new_val = $w + 1$ values from $N(\mu(\text{neigh}), \sigma^2(\text{neigh}))$;
    $i'[t - \frac{w}{2}, t + \frac{w}{2}]$ = new_val;
    act_diff$[t] = \phi(i)[t] - \phi(i')[t]$;
  end
  return act_diff;

Synthetic limited-duration sinusoidal burst signals $s_{f,d,a}(t)$ as in Eq. 1 were then used as network inputs to better characterize the time-frequency response of particular filters of interest. Heat maps were employed to display the resulting internal activation levels. These heat maps were coordinatized by the frequency, duration, and amplitude parameters of the input signal – a representation that is relevant to exploring the response to sleep spindles, alpha wave bursts, and slow-wave sleep. Note that the activation of each filter (unit), ϕ, in response to a particular input signal, $s = s(t)$, is a time series $\phi(s) = \phi(s)(t)$. The maximum activation for times t in the middle epoch of each five-epoch input sample is used to generate the heat map. See Algorithm 3. Two-dimensional projections of the heat maps were used for presentation purposes, as three-dimensional heat maps are difficult to interpret on printed paper without the possibility of direct exploratory interaction.

Algorithm 3: Heat map generation procedure

Function *generate heat map(filter of interest, ϕ)*
  forall *frequencies f in* $[0.5, 20]$ *Hz (0.5 Hz steps)* **do**
    forall *durations d in* $[0.5, 6]$*s (0.5s steps)* **do**
      forall *amplitudes a in* $[20, 330]$ μV *(10 μV steps)* **do**
        use synthetic burst $s_{f,d,a}(t)$ as input to net,
          centered in five-epoch input window,
          extended with zeros everywhere else;
        $r_\phi(f, d, a) = \max_t \phi(t)$ for t in middle epoch;
      end
    end
  end
  return heat map of responses $r_\phi(f, d, a)$;

A visualization tool based on [Yosinski et al. 2015] was developed to facilitate the inspection of internal network activations, both across an entire layer, as well as for individual filters. See Fig. 5.

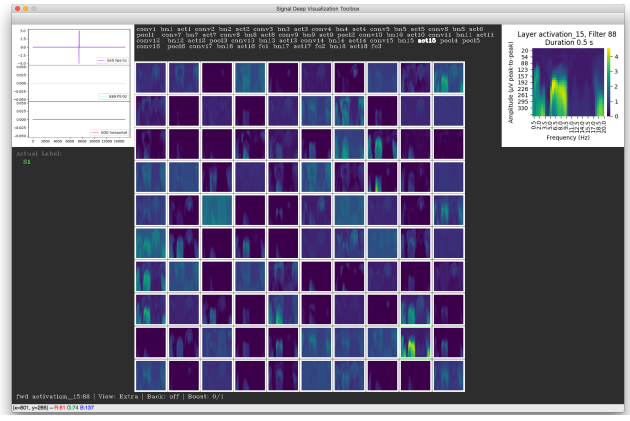

Figure 5: Sample screen shot of the visualization tool developed for the present work. This view shows activation heat maps for the filters in one of the layers as in Algorithm 3, with a close-up for a selected filter at top right. Alternate views are available that show filter activations as functions of time, the stage distribution of a layer as in Algorithm 1, and the effect of user-defined input transformations, such as in the activation difference analysis of Algorithm 2.

4 RESULTS AND DISCUSSION

4.1 Classification performance

A summary of the predictive performance of the CNN of Fig. 3 appears in Table 1. Our model attains an overall classification accuracy of 81%, outperforming another CNN approach [Tsinalis et al. 2016a] on the same data set, which reached an overall accuracy of 74% with training on one EEG signal using a shallower, six-layer architecture. After modifying our model to train on one signal only and evaluating performance on each fold, total accuracy decreased from 81% to 80%, still handily outperforming [Tsinalis et al. 2016a]. This shows that greater network depth has a greater effect on performance than does the use of additional signal channels.

Notice in Table 1 that performance differs among sleep stages: stages S2, S3, and REM are modeled much better by the network than are Wake and S1. This difference is likely tied to the fact that Wake and S1 are comparatively under-represented in the data set.

Table 1: Cumulative confusion matrix with precision, recall, and F1-score [Sokolovsky et al. 2018]. Table entries correspond to the number of epochs belonging to a given class as classified by a technician (rows) and by the model (columns). Overall accuracy is 81%**, as compared with** 82.6% **agreement among expert human scorers [Rosenberg et al. 2013].**

	S1	S2	S3	Wake	REM	Recall	F1
S1	1299	595	10	382	406	48%	46%
S2	634	15402	691	438	572	87%	87%
S3	4	654	4972	71	1	87%	87%
A	666	178	21	1441	128	59%	57%
R	373	719	4	260	6348	82%	84%
Precision	44%	88%	87%	56%	85%		

Our model's overall classification accuracy of 81% compares favorably with the gold standard that is the level of agreement among expert human scorers, 82.6% [Rosenberg et al. 2013]. Recent results on the same dataset [Supratak et al. 2017] further suggest that our model extracts approximately as much sleep stage information from EEG data as human experts. [Supratak et al. 2017] reports a maximum accuracy of 82%, only a 1% improvement over our results, using a more complicated CNN and bi-directional LSTM model.

4.2 Sleep stage specialization among filters

The collection of sleep stage distributions associated with the top-activating signals for the filters in a given layer as in Algorithm 1, is one view of that layer's learned behavior. Filters in "shallow" layers – those close to the input – generally have sleep stage distributions that show little preference for specific stages. Specialization of filters by sleep stage emerges in deeper layers. For example, while only one filter in layer 1 (shallow) satisfies the condition that 75% or more of its top activating signals belong to the same sleep stage, well over half of the filters in layer 13 (deep) satisfy this stage specialization condition. The deepest layers include filters that specialize in stages S2, S3, and REM – precisely the stages for which predictive performance is best as shown in Table 1. No filters were found that specialize in stages Wake and S1, which are the two stages for which the least amount of training data were available (compare row sums in Table 1).

4.3 Emergence of internal features

The trend toward greater sleep stage specialization with depth levels off a few layers before the deepest layer. This fact, combined with the fact that greater network depth does lead to improved classification performance, as discussed above, is tied to the fact that the filters of the network learn more detailed information than stage labels alone. We will describe several of the internal features that emerge during learning, showing that they correspond to specific features that figure prominently in the AASM staging standard [Berry et al. 2012], which expert human scorers rely on closely. We will also show that deeper, more complex features are built from simpler features that emerge in shallower layers, thus demonstrating a phenomenon known to occur in CNN for two-dimensional image classification [Zeiler and Fergus 2014], but that not has previously been reported for one-dimensional sleep EEG.

4.3.1 The impact of depth. Overall network depth (number of layers) is important in determining CNN predictive performance, with greater depth yielding higher classification accuracy [Sokolovsky et al. 2018]. The results of the present paper show that this phenomenon can be traced back to the complexity of the features that are learned by layers at different depths. Shallower layers tend to capture only simple features. Specifically, we find that many filters in the shallowest layers serve as narrow-band frequency detectors. Fig. 6 shows three examples, corresponding to filters that respond to selected frequencies in the delta (0.5 − 2 Hz, left), sigma (11 − 16 Hz, middle), and beta (16+ Hz, right) ranges. These particular frequency ranges are associated, respectively, with slow waves in S3 sleep, spindles in S2 sleep, and ideation in wakefulness and REM sleep. These shallower filters do not differentiate among input signal amplitudes or durations. We will show, below, that, by refining and combining the elements identified in the shallower filters, such as specific EEG frequency bands, filters in deeper layers capture more complex relationships among frequency, amplitude, and duration. See Fig. 7 and the discussion below. It is this richer feature vocabulary that emerges in the deeper layers that enables human-expert-level classification performance.

4.3.2 Slow-wave sleep (SWS). An important emergent feature relates to stage S3 (deep) 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 [Berry et al. 2018]). The frequency range associated with slow-wave sleep is identified in shallower layers of the network, without yet being combined with the requisite amplitude information. Fig. 8 provides an example of the response to a data sample of actual stage S3 human sleep, of a particular filter in one of the deeper convolutional layers of the network described in section 3.1. A visualization technique based on [Zintgraf et al. 2017] 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 greatest impact in the

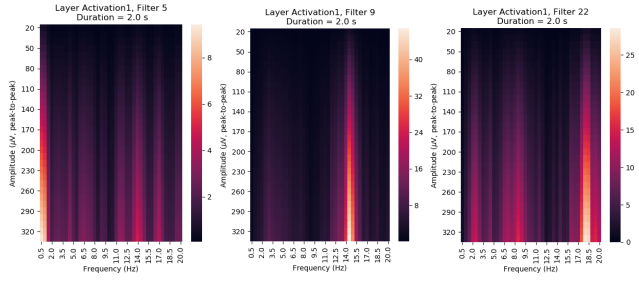

Figure 6: Several filters in the shallowest convolutional layer respond mainly within a single narrow frequency band. In this and subsequent figures, each heat map represents a filter's response to synthetic bursts of various frequencies (horizontal axis) and amplitudes (vertical axis). The three filters shown respond to selected frequencies in the delta (0.5 – 2 Hz, left), sigma (11 – 16 Hz, middle), and beta (16+ Hz, right) ranges.

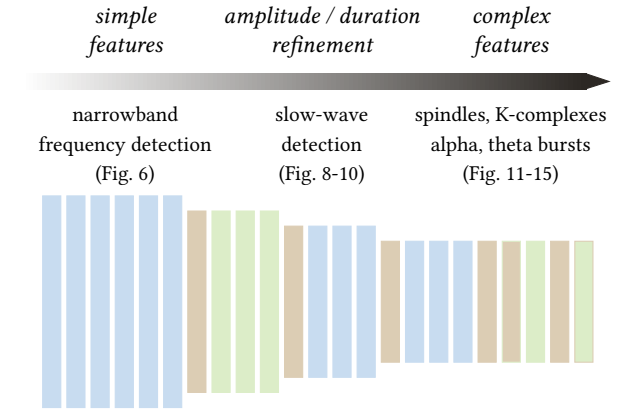

Figure 7: Emergence of sleep EEG features with increasing depth of the convolutional layers within the CNN of Fig. 3. Simple features first emerge in shallow layers, and are refined and combined into complex features in deeper layers.

activation of this filter is produced by the contiguous segment of slow waves toward the center of the EEG plot.

It is worth noting that the filter's response occurs at far finer time scales – two or three seconds – than the 30s epoch resolution of the available supervised training information for the network, namely the sleep stage labels of the input signals. The use of synthetic input data provides more detailed information about the frequency, duration, and amplitude response of this and similar filters. Indeed, the heat maps in Fig. 9 show that the response to synthetic bursts of a similar slow-wave detecting filter in our network is greatest in the frequency range of approximately 0.5 – 2 Hz, for peak-to-peak signal amplitudes above 80μV or so, and burst durations of at least 2s. The response characteristics of this particular filter closely match the AASM criteria for slow waves in stage S3 sleep. Fig. 10 shows that the SWS feature is refined with increasing depth of the convolutional layers – a pattern that is repeated for other emergent features. See Fig. 7.

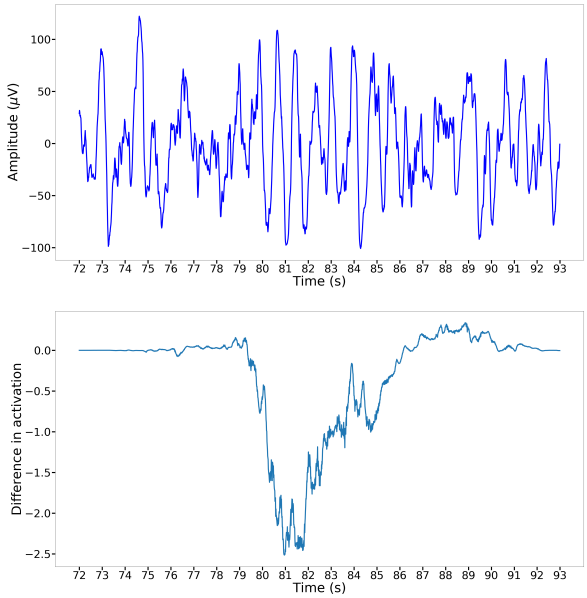

Figure 8: Response to stage S3 human sleep of a convolutional filter in our deep CNN that learns to detect slow waves. Top plot shows one EEG input channel. Bottom plot shows the difference in filter activation associated with removing a portion of the input at the corresponding locations along the time axis. The slow wave train toward the center has the greatest effect on activation.

4.3.3 S2 sleep. Certain convolutional filters in deeper network layers respond to the brief (0.5 – 2*s*) EEG signal bursts at 11 – 16 Hz known as sleep spindles, as well as to the high-amplitude sawtooth-like K-complexes. Both of these features are characteristic of sleep stage S2, as described in the AASM standard. Fig. 11 shows the response of one such filter in layer 13 to a sample of actual stage S2 human sleep that contains a sleep spindle and a K-complex.

The heat maps in Fig. 12 show that this filter exhibits a high response to synthetic sinusoidal wave bursts of frequencies (11 – 16 Hz) and durations (0.5 – 1.5s) that are consistent with sleep spindles. The heat maps reveal that this filter also responds to the low-frequency, high-amplitude waves that characterize slow-wave sleep. The slow-wave response in Fig. 12 is very similar to that of the SWS filter from a shallower layer in Fig. 9, though the former, deeper filter has developed a more refined preferred range of high amplitudes (80 – 250μV or so in Fig. 12, as compared with simply 80+ μV in Fig. 9). This illustrates a recurring theme: the use of simpler features from shallower layers as building blocks for more complex features in deeper layers. Likewise, Fig. 13 shows that the general pattern of refinement of features with increasing depth applies to the S2 spindle / K-complex feature, in particular.

No filters were found that respond to sleep spindles alone, without some degree of response in the slow wave range also. This phenomenon may be understood in terms of the optimization process involved in supervised learning, together with the low-frequency content of K-complexes. Because sleep spindles involve a relatively

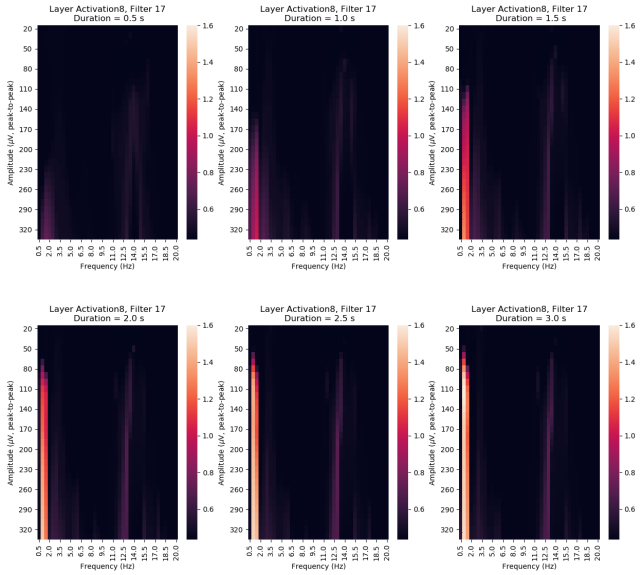

Figure 9: Frequency-amplitude response to different synthetic burst durations, of a filter in a mid-depth convolutional layer that learns to detect SWS. Burst duration increases from top left to bottom right. Response to SWS increases with burst duration, and is greatest for durations of 2s or more, and amplitudes of at least 80μV (bottom row, left edge of each map).

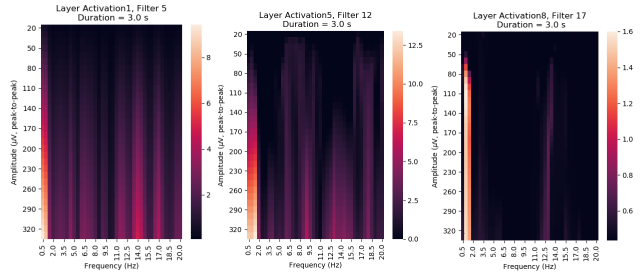

Figure 10: Refinement of S3 SWS feature with increasing depth (left to right). Response to frequencies below 1 Hz arises in the shallowest layer (left). Filters in slightly deeper layers (center and right) develop more focused responses in frequency ($0.5 - 2$ Hz) and amplitude (80μV or greater), as well as duration (not shown), thus capturing key defining characteristics of slow waves, even at moderate depths.

complex relationship among frequency, duration, and amplitude, filters that capture sleep spindles well can only occur in the deeper layers of the network. However, such deep layers are relatively close to the output, where the objective of minimizing sleep stage classification error exerts a strong pressure to model individual sleep stages well. As a result, filters in the deeper layers tend to be associated with specific sleep stages. Large-amplitude slow waves are the dominant feature of S3 sleep, and S3 contains minimal sigma-band power (the spindle range). In contrast, S2 sleep contains a

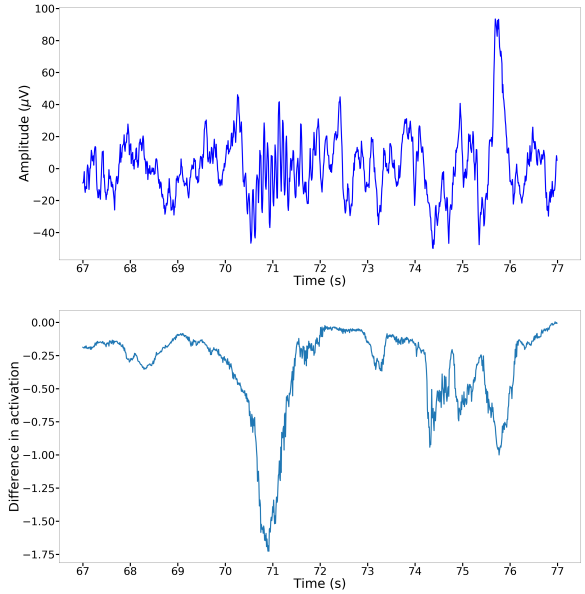

Figure 11: Response to stage S2 human sleep of a convolutional filter in our deep CNN that learns to detect sleep spindles and K-complexes. Top plot shows one EEG input channel. Bottom plot shows the difference in filter activation associated with removing a portion of the input at the corresponding locations along the time axis. The spindle toward the center has the greatest effect on activation, followed by the K-complex at far right.

substantial amount of delta-band (slow-wave) power [Armitage 1995] that is associated with the K-complexes that often accompany sleep spindles [Amzica and Steriade 1997]. These facts lead to slow-wave-based filters for stage S3 as in Fig. 9, and to mixed spindle-plus-slow-wave filters for stage S2 as in Fig. 12.

4.3.4 REM sleep. Certain convolutional filters respond mainly to REM sleep. The EOG channel can be expected to be especially relevant, as it provides measurements associated directly with the eye movements that give REM its name. This expectation is confirmed by our results. Nonetheless, filters that capture EEG aspects of REM sleep were also found. For example, the EEG response heat map in Fig. 14 of a particular filter in the deepest convolutional layer shows special sensitivity to moderate-amplitude alpha-range waves of duration $2 - 3s$ that are observed in REM sleep [Cantero and Atienza 2000]. Additional response to EEG in the theta and beta ranges rounds out a spectral profile characteristic of REM [Armitage 1995]. The REM feature is refined with depth, as occurs for the SWS and S2 features discussed previously. See Fig. 15.

4.3.5 Wakefulness and S1 sleep. None of the convolutional filters in the trained CNN respond primarily to Wake or S1 sleep stages. In fact, very few filters respond substantially to input signals of either of these two stages. Among filters with a moderate S1

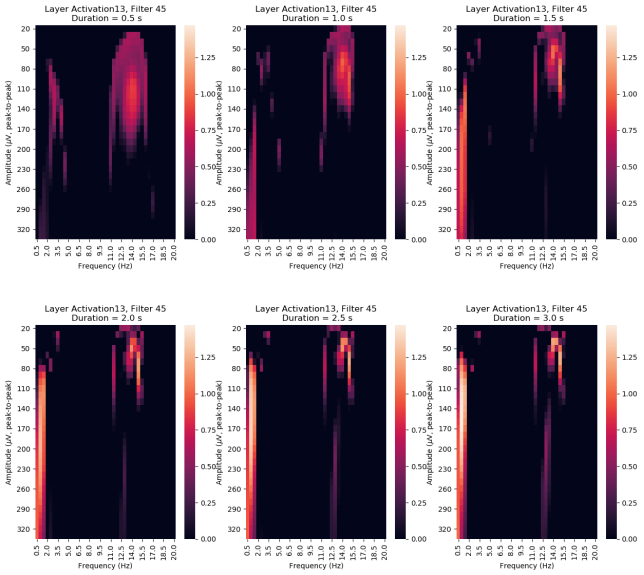

Figure 12: Frequency-amplitude response to different synthetic burst durations, of a filter in a deeper convolutional layer that learns a combination of spindle and SWS features. Burst duration increases from top left to bottom right. Response to spindles shows more structure for durations of 1.5s or less (top row, upper right of each map), while SWS response is most prominent for durations above 2s (bottom row, left edge of each map).

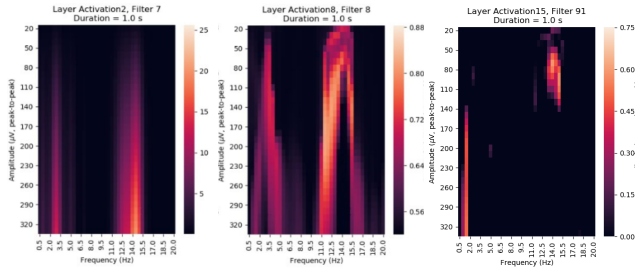

Figure 13: Emergence of S2 / S3 sleep spindle feature with increasing depth. Filters in shallow layers (left) respond to the characteristic $11-15$ Hz spectral range, without differentiation by amplitude or duration. Deeper layers (center and right) develop a more focused response in frequency and amplitude, as well as duration (not shown), thus capturing all defining characteristics of sleep spindles.

response, a common feature appears to be a sensitivity to shorter-duration, lower-amplitude bursts in the theta ($4-7$ Hz) and alpha ($8-12$ Hz) spectral bands, as well as the beta band (16+ Hz). Such filters also respond to REM (also a lighter sleep stage) more so than to Wake episodes. Qualitative appearance of the response heat maps of such filters (not shown) is similar to the REM filters in Fig. 14 (0.5s duration, top left), but with greater beta content. The lack of a more distinctive heat map signature is consistent with the relative

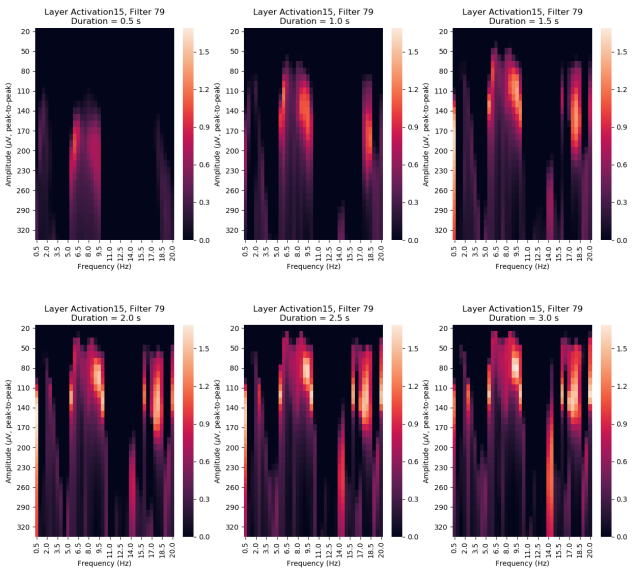

Figure 14: Frequency-amplitude response to different synthetic burst durations, of a filter in a deeper convolutional layer that responds primarily to REM sleep. Burst duration increases from top left to bottom right. The greatest response occurs for lower-amplitude signals of duration 2s or greater (bottom row) in the lower alpha range ($8-10$ Hz, upper middle of each map), with additional theta ($5-7$ Hz, left of alpha) and beta (16+ Hz, right end) response.

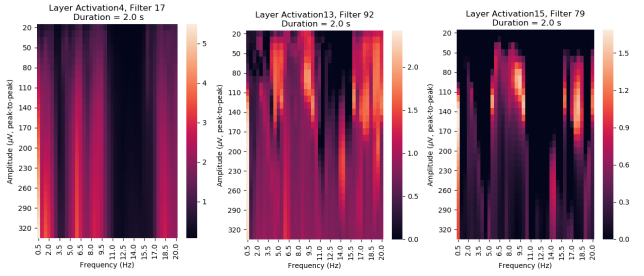

Figure 15: Emergence of REM features with increasing depth (left to right). Shallow layers (left) show response to theta ($5-7$ Hz), lower alpha ($8-10$ Hz), beta ($17-20$ Hz) and delta ($0.5-2.5$ Hz) ranges, without differentiation by amplitude. Filters in deeper layers (center and right) develop a more focused response to smaller amplitudes in the alpha range (with preferential duration 2s or greater, not shown), and to moderate amplitudes in the other ranges, thus capturing key EEG characteristics of REM sleep.

scarcity of S1 and Wake samples in the training data, leading to comparatively weak modeling relative to other sleep stages. This issue merits greater attention in future work.

5 CONCLUSIONS

We described our work in predictive modeling in sleep medicine using a deep convolutional neural network (CNN) architecture for sleep stage classification of multi-channel polysomnogram (PSG) data, together with domain interpretation of the emergent internal features. The classification performance of the proposed CNN model is at approximately the same level as the rate of agreement among human expert scorers, which can reasonably be taken as an upper bound on objectively meaningful performance in this context. We showed that the excellent performance of this model is due mostly to its depth (17 convolutional layers, with additional pooling and batch normalization layers), and not to the use of multi-channel data, as using a single EEG channel results in only a minor decrease in classification accuracy (80% vs. 81%).

Using visualization and analysis of network response to natural human sleep data and synthetic burst data, we identified specific emergent EEG features in the convolutional filters of the CNN, including large-amplitude slow waves, sleep spindles and K-complexes, and low-amplitude alpha waves, that human experts rely on for the visual scoring of sleep. The complexity of the emergent features increases with depth within the CNN. Many filters in shallow layers act as narrowband frequency detectors. Deeper layers refine and combine the more elementary features to capture patterns that are representative of individual sleep stages. For example, an emerging slow-wave feature is refined in deeper layers to account for a preferred range of amplitudes. Other convolutional filters in deep layers respond sensitively to both sleep spindles and K-complexes – features that co-occur in sleep stage S2. Yet other filters respond to the alpha bursts that occur in REM sleep. This paper provides an important new demonstration of hierarchical feature emergence in CNN for physiological time series data, the first such work that we are aware of in sleep stage classification.

ACKNOWLEDGMENTS

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

REFERENCES

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. 2016. TensorFlow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467* (2016).

Florin Amzica and Mircea Steriade. 1997. The K-complex: Its slow (< 1-Hz) rhythmicity and relation to delta waves. *Neurology* 49, 4 (1997), 952–959. https://doi.org/10.1212/WNL.49.4.952 arXiv:http://n.neurology.org/content/49/4/952.full.pdf

R. Armitage. 1995. The distribution of EEG frequencies in REM and NREM sleep stages in healthy young adults. *Sleep* 18, 5 (June 1995), 334–341.

Richard B Berry, Rita Brooks, Charlene E Gamaldo, Susan M Harding, CL Marcus, and BV Vaughn. 2012. The AASM manual for the scoring of sleep and associated events. Rules, Terminology and Technical Specifications. *American Academy of Sleep Medicine* (2012).

Richard B Berry et al. 2018. (Updates to) The AASM Manual for the Scoring of Sleep and Associated Events. (2018). https://aasm.org/clinical-resources/scoring-manual/

José Luis Cantero and Mercedes Atienza. 2000. Alpha burst activity during human REM sleep: descriptive study and functional hypotheses. *Clinical Neurophysiology* 111, 5 (2000), 909 – 915. https://doi.org/10.1016/S1388-2457(99)00318-1

Xudong Cao. [n. d.]. A practical theory for designing very deep convolutional neural networks. ([n. d.]). https://kaggle2.blob.core.windows.net/forum-message-attachments/69182/2287/A%20practical%20theory%20for%20designing%20very%20deep%20convolutional%20neural%20networks.pdf

François Chollet et al. 2015. Keras. https://keras.io. (2015).

Sander Dieleman and Benjamin Schrauwen. 2014. End-to-end learning for music audio. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 6964–6968.

D. Gabor. 1946. Theory of communication. Part 1: The analysis of information. *Electrical Engineers - Part III: Radio and Communication Engineering, Journal of the Institution of* 93, 26 (November 1946), 429–441. https://doi.org/10.1049/ji-3-2.1946.0074

AL Goldberger, LAN Amaral, L Glass, JM Hausdorff, PC Ivanov, RG Mark, JE Mietus, GB Moody, C-K Peng, and HE Stanley. 2000. Physiobank, physiotoolkit, and physionet. *Circulation* 101, 23 (2000), e215–e220.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

Madeleine M Grigg-Damberger. 2009. The AASM scoring manual: a critical appraisal. *Current opinion in pulmonary medicine* 15, 6 (2009), 540–549.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*. 1026–1034.

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167* (2015).

Eric Kauderer-Abrams. 2018. Quantifying Translation-Invariance in Convolutional Neural Networks. *CoRR* abs/1801.01450 (2018). arXiv:1801.01450 http://arxiv.org/abs/1801.01450

B Kemp, AH Zwinderman, B Tuk, HAC Kamphuisen, and JJL Oberyé. 2000. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the EEG. *IEEE-BME* 47, 9 (2000), 1185–1194.

Serkan Kiranyaz, Turker Ince, and Moncef Gabbouj. 2016. Real-time patient-specific ECG classification by 1-D convolutional neural networks. *IEEE Transactions on Biomedical Engineering* 63, 3 (2016), 664–675.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

Martin Längkvist, Lars Karlsson, and Amy Loutfi. 2012. Sleep stage classification using unsupervised feature learning. *Advances in Artificial Neural Systems* 2012 (2012). https://doi.org/doi:10.1155/2012/107046

G. Medic, M. Wille, and M. E. Hemels. 2017. Short- and long-term health consequences of sleep disruption. *Nature and Science of Sleep* 9 (2017), 151–161. https://doi.org/10.2147/NSS.S134864

Karol J Piczak. 2015. Environmental sound classification with convolutional neural networks. In *Machine Learning for Signal Processing (MLSP), 2015 IEEE 25th International Workshop on*. IEEE, 1–6.

Waseem Rawat and Zenghui Wang. 2017. Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. *Neural Comput.* 29, 9 (Sept. 2017), 2352–2449. https://doi.org/10.1162/neco_a_00990

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*. 91–99.

Marco Roessen and Bob Kemp. 2018. Polyman EDF+ viewer. (2018). http://www.edfplus.info

Richard S Rosenberg, Steven Van Hout, et al. 2013. The American Academy of Sleep Medicine inter-scorer reliability program: sleep stage scoring. *J Clin Sleep Med* 9, 1 (2013), 81–87.

Michael H Silber, Sonia Ancoli-Israel, Michael H Bonnet, Sudhansu Chokroverty, Madeleine M Grigg-Damberger, Max Hirshkowitz, Sheldon Kapen, Sharon A Keenan, Meir H Kryger, Thomas Penzel, et al. 2007. The visual scoring of sleep in adults. *J Clin Sleep Med* 3, 2 (2007), 121–131.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

M. Sokolovsky, F. Guerrero, S. Paisarnsrisomsuk, C. Ruiz, and S.A. Alvarez. 2018. Human expert-level automated sleep stage prediction and feature discovery by deep convolutional neural networks. *BIOKDD 2018* (2018).

Akara Supratak, Hao Dong, Chao Wu, and Yike Guo. 2017. DeepSleepNet: a Model for Automatic Sleep Stage Scoring based on Raw Single-Channel EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 25 (2017), 1998–2008.

Orestis Tsinalis, Paul M Matthews, and Yike Guo. 2016a. Automatic sleep stage scoring using time-frequency analysis and stacked sparse autoencoders. *Annals of biomedical engineering* 44, 5 (2016), 1587–1597.

Orestis Tsinalis, Paul M Matthews, Yike Guo, and Stefanos Zafeiriou. 2016b. Automatic Sleep Stage Scoring with Single-Channel EEG Using Convolutional Neural Networks. *arXiv preprint arXiv:1610.01683* (2016).

Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. 2015. Understanding Neural Networks Through Deep Visualization. In *Deep Learning Workshop, International Conference on Machine Learning (ICML)*.

Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision (ECCV)*. Springer, 818–833.

Haomin Zhang, Ian McLoughlin, and Yan Song. 2015. Robust sound event recognition using convolutional neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 559–563.

Luisa M. Zintgraf, Taco S. Cohen, Tameem Adel, and Max Welling. 2017. Visualizing Deep Neural Network Decisions: Prediction Difference Analysis. *CoRR* abs/1702.04595 (2017). arXiv:1702.04595 http://arxiv.org/abs/1702.04595