IOP Publishing Physiol. Meas. 40 (2019) 074001 (18pp) https://doi.org/10.1088/1361-6579/ab2664

Physiological Measurement

IPEM Institute of Physics and Engineering in Medicine

PAPER

Hybrid scattering-LSTM networks for automated detection of sleep arousals

RECEIVED
1 March 2019

REVISED
28 May 2019

ACCEPTED FOR PUBLICATION
3 June 2019

PUBLISHED
19 July 2019

Philip A. Warrick[1], Vincent Lostanlen[2] and Masun Nabhan Homsi[3]

[1] PeriGen. Inc., Montreal, Canada
[2] New York University, New York, NY, United States of America
[3] Simón Bolívar University, Caracas, Venezuela

E-mail: philip.warrick@perigen.com, vincent.lostanlen@nyu.edu and mnabhan@usb.ve

Keywords: polysomnography, sleep wake disorders, arousal, scattering transform, convolutional neural networks, long short term memory networks

Abstract

Objective: Early detection of sleep arousal in polysomnographic (PSG) signals is crucial for monitoring or diagnosing sleep disorders and reducing the risk of further complications, including heart disease and blood pressure fluctuations. *Approach*: In this paper, we present a new automatic detector of non-apnea arousal regions in multichannel PSG recordings. This detector cascades four different modules: a second-order scattering transform (ST) with Morlet wavelets; depthwise-separable convolutional layers; bidirectional long short-term memory (BiLSTM) layers; and dense layers. While the first two are shared across all channels, the latter two operate in a multichannel formulation. Following a deep learning paradigm, the whole architecture is trained in an end-to-end fashion in order to optimize two objectives: the detection of arousal onset and offset, and the classification of the type of arousal. *Main results and Significance*: The novelty of the approach is three-fold: it is the first use of a hybrid ST-BiLSTM network with biomedical signals; it captures frequency information lower (0.1 Hz) than the detection sampling rate (0.5 Hz); and it requires no explicit mechanism to overcome class imbalance in the data. In the follow-up phase of the 2018 PhysioNet/CinC Challenge the proposed architecture achieved a state-of-the-art area under the precision-recall curve (AUPRC) of 0.50 on the hidden test data, tied for the second-highest official result overall.

1. Introduction

Sleep induces many physiological correlates in the brain, which vary consistently through time and across subjects. Through time, they delineate a sequence of sleep stages, from wakefulness to deep sleep, as well as so-called 'paradoxical' rapid eye motion (REM) sleep. Across subjects, they are harbingers of the presence of sleep-related disorders. There is a growing amount of evidence that such disorders are linked with multiple widespread pathologies, including weight gain, depression, heart disease, and diabetes. Consequently, advancing our current knowledge of the neurophysiology of sleep in humans is, more than a challenging research question, also an issue of public health.

The recurrence of short interruptions during sleep, known as *arousals*, are of particular concern to medical practitioners (Halász *et al* 2004). This is because they reflect a broad range of symptoms, among which sleep apnea is the most common, yet certainly not the only possible one causing disturbance of sleep. Sleep arousals can also be spontaneous, result from teeth grinding, partial airway obstructions, or even snoring (Ghassemi *et al* 2018a) and it is these types of non-apneal arousals that are the focus of our paper. Furthermore, arousals cause a sudden change from REM sleep to non-REM sleep, and may in some cases cause prolonged wakefulness. Therefore, within the more general scope of improving the quality of sleep, the focus on accurately and precisely detecting sleep arousals plays a key role.

Inside the brain, arousal disrupts the patterns of electrical activity that are typical of healthy sleep. Electroencephalography (EEG) can directly measure the resulting disruption. Besides EEG, since arousals affect the entire body, they indirectly impact electromyogram (EMG), electrooculogram (EOG), electrocardiogram (ECG), oxygen

saturation (SaO_2), respiratory airflow (AIRFLOW) and respiratory movements (CHEST and ABD). The collection and gathering of all these aforementioned vital signs under one multidimensional time series is a clinical procedure that is known as polysomnography (PSG). Certified sleep technologists have learned the skill of visually interpreting the temporal oscillations of these PSG recordings, so as to pinpoint the presence of sleep arousals.

One major hindrance to the scalability of diagnosing sleep disorders from PSG recordings is the amount of human labor that their annotation incurs. Indeed, given that sleep arousals only last for a few seconds at once, they appear on PSG as transient phenomena, hence requiring practitioners to scan whole signals at a relatively fine temporal resolution. This shortcoming calls for the facilitation, if not the complete automation, of sleep arousal monitoring. In the future, a fast and accurate computational system for detecting sleep arousals could potentially be shipped with PSG hardware toolkits, so that timestamps of predicted arousals would appear on screen as a supplement to raw PSG acquisition.

The physiological frequency bands of interest for the PSG signals generally range from 0.05 to several hundreds of cycles per second. The lowest (delta) band of conventional EEG study has a lower limit of 0.5 Hz or 1.0 Hz (depending on the definition) while 100 Hz corresponds to the highest frequency of the EEG gamma band (Niedermeyer and da Silva 2004). EEG frequency bands as low as 0.1 and 0.01 Hz have also been identified as potentially informative, referred to as 'slow-frequency' and 'infra-slow' EEG bands, respectively (Hiltunen *et al* 2014). The ECG frequency spectrum is generally considered to be 0.05–100 Hz (Clifford *et al* 2006), although Jarvis and Mitra (2000) reported ECG activity related to sleep apnea down to 0.02 Hz. EMG ranges from 5.0 Hz to much higher frequencies up to 450 Hz (Viitasalo and Komi 1977). The other modalities of respiratory movements, AIRFLOW and SaO_2 are lower frequency phenomena whose respective coherence spectra show activity between 0.05 and 0.35 Hz (Nino *et al* 2013).

In this paper, we present a new computational system for the detection and classification of sleep arousals from PSG recordings. The main contribution of this paper is to demonstrate that it is practically feasible, and even beneficial to inter-subject generalization, to integrate many different physiological time scales and modalities into a single time-invariant model for PSG monitoring. First, our system is *multiscale*: it combines ST coefficients with convolutional operators, thus extracting information at time scales ranging from the millisecond to the minute, both in terms of baseband and narrowband amplitude modulations of power spectral density. Secondly, our system is *multimodal*: rather than engineering a different transformation to each one of the modalities of polysomnography, it learns a single transformation via the convolutions that is shared across all of them: EEG, ECG, EMG, and so forth. We find that inducing an autoregressive structure that is both multiscale and multimodal allows us to train a time-invariant deep learning model whose effective receptive field is considerably larger than in a linear autoregressive model. Indeed, each recursive update of latent variables in the model relies on the values of over 10^5 samples in the multidimensional time domain of PSG signals; yet, the model remains computationally tractable and incurs little statistical overfitting, because observations in the distant past affect comparably less trainable parameters than samples in the recent past.

Our system composes three recently published methods in signal processing and deep learning: scattering transform (ST); depthwise-separable convolutional layers (CNN); and bidirectional long-short term memory (BiLSTM) layers. Although these methods may have found biomedical engineering applications in the past, this paper is the first to implement all of them together, as parts of a common end-to-end pipeline. In particular, to the best of our knowledge, no previously published machine-learning system has employed the scattering transform as a multimodal frontend to either, let alone both, depthwise-separable or bidirectional models.

We report numerical results of our best performing model, as well as of some less sophisticated variants, upon participating in the 'You Snooze, You Win' PhysioNet/CinC 2018 Challenge for automatic detection and classification of sleep arousals (Ghassemi *et al* 2018a). The dataset of this challenge contains 994 time series for training and 989 time series for testing. This paper expands upon our work submitted to the 2018 Challenge (Warrick and Homsi 2018b) which ranked sixth in the official phase. Herein we discuss the methodological implications of having adopted a multiscale and multimodal approach to the problem and propose three further architectural improvements to both accuracy and speed. First, it is the first use of a hybrid scattering transform-bidirectional long short term memory (ST-BiLSTM) network with biomedical signals. Secondly, it captures frequency information much lower (0.1 Hz) than the detection sampling rate (0.5 Hz). Thirdly, we notice that, in the latest version of our model, it is no longer necessary to explicitly counteract the class imbalance of the training data by way of an importance sampling heuristic.

2. Related work

In this section, we briefly review the prior literature on computational methods for sleep arousal detection. Our review consists of two parts, respectively corresponding to methods outside and inside the 2018 Challenge.

2.1. Prior approaches

Historically, most published methods for detecting transients in biomedical signals are composed of two stages: domain-specific feature extraction and general-purpose machine learning. In the context of polysomnography, features are engineered either in the time domain, the frequency domain, or the time-scale domains. Furthermore, the most widespread machine learning methods are decision trees (De Carli *et al* 1999, Agarwal 2006, Shmiel *et al* 2009), multi-layer perceptrons (MLP) (Huupponen *et al* 1996), and support vector machines (SVM) (Cho *et al* 2006).

The design of handcrafted features and finding their optimal combination for improving classifier performance can be difficult and time-consuming; to overcome this issue in recent years, the feature-engineering step is often automated using deep neural networks (DNN) such as CNNs (Tsinalis *et al* 2016, Aggarwal *et al* 2018, Zhang and Wu 2018).

De Carli *et al* (1999) recorded EEG (F4-C4 and C4-O2 channels) and a chin EMG to develop their arousal detector, using the American Academy of Sleep Medicine criteria defining all types of arousal (AASM 1992). Their dataset was relatively small, consisting of eleven overnight recordings of patients with various pathological conditions. They established their ground truth from an aggregation of two human experts. Then, they compared the arousal detection of their linear-discriminant classifier to the markings of the same two human experts on independent test data. They found that the sensitivity was higher (88.1% versus 72.4% and 78.4%) while the precision was lower (74.5% versus 83.0% and 82.0%). Since ground truth was derived from the marking of these same humans, the reported human scores are possibly too optimistic to reflect a real-world use case. Despite this caveat and the small dataset size, the results of De Carli *et al* (1999) offer a useful insight that automating the detection of sleep arousals has great application potential, yet requires the development of advanced techniques in signal processing and machine learning.

Although numerous publications have proposed to apply deep learning to sleep stage classification (wakefulness, stage 1, stage 2, stage 3 and REM) (Tsinalis *et al* 2016, Aggarwal *et al* 2018, Zhang and Wu 2018), fewer works have applied them to sleep arousal data. Recently, Saeed *et al* (2017) explored the potential of CNN for classifying sleep arousals into three categories: under-aroused, normal, and over-aroused. They measured raw physiological signals collected from wrist-wearable devices, as worn by eleven subjects. They found that their CNN outperformed a non-convolutional baseline, with respective F-scores of 0.82 and 0.75.

2.2. Approaches of the 2018 Challenge

Our approach submitted to the 2018 Challenge was rated sixth out of 19 teams during the official phase. These top-six approaches all used DNNs, and four of these, including ours, also used an ensemble method. Therefore, this section is divided in two subsections: the first one concerns approaches that used a single deep model, while the second focuses on those approaches that used an ensemble method. All performance measures refer to the area under the precision-recall curve (AUPRC) of the algorithm on the hidden test set. Table 1 shows the various studies conducted on the automated detection of arousal regions in PSG signals, along with the results yielded by the present study which are discussed in section 4.

2.2.1. Deep architecture

He *et al* (2018) first partitioned PSG signals into smaller segments of 100 s to overcome the class-imbalance problem and to reduce the long training time required for the large dataset. The segments were then fed into a sequence-to-sequence neural network composed of two parts. The first part was dedicated to feature extraction which consisted of a 1D-CNN and a BiLSTM, while the second part was for feature classification which was composed of a fully connected layer. The proposed NN could effectively identify the arousal regions with an AUPRC of 0.43.

Varga *et al* (2018) combined hand-crafted polysomnographic features with a deep-learning architecture. Sixty-eight features from time, frequency, statistical and information-theoretic domains were extracted from EEG, EOG, EMG, AIRFLOW and ECG signals, while a scaled version of SaO_2 was used directly. The extracted features were first normalized and then resampled at 21 times over a two-minute window, with denser sampling around the central current feature sample. The resulting 68×21 feature map was advanced at 1 s intervals. The DNN consisted of a 2D convolution layer and two dense layers with seven outputs; two of which were for arousal and non-arousal classification, while the remaining ones were for sleep stage classification. Class imbalance was addressed by discarding all invalid arousal samples and some non-arousal samples, ensuring that at least 25% of the training feature samples had the (minority) arousal target. Other experiments used the sleep stage as an auxiliary target with dropout and post-processing strategies, but the performance of the proposed NN did not significantly improve. This approach achieved an AUPRC of 0.42.

Miller *et al* (2018) used a convolutional-deconvolutional neural network inspired by densely connected CNN and semantic segmentation networks (Jégou *et al* 2017). Their model consisted of eight convolutional layers, eight deconvolutional layers, and a dense-softmax layer to calculate the probability distribution over the

Table 1. Comparison of DNN-based 2018 Challenge approaches.

Reference	Excluded channels	Preprocessing	Base classifier	Ensemble classifier	AU-PRC
Howe-Patterson *et al* (2018)	ECG	FIR, down-sampled, FFT convolution. Scaling SaO_2	Multiple dense convolutional units + BiLSTM with four outputs	Averaging per sample over four fold-classifiers	0.54
Þráinsson *et al* (2018)	NA	Time, frequency and statistical.	BRNN-LSTM and a dense layer with 50 outputs	Averaging per sample over five fold-classifiers	0.45
He *et al* (2018)	NA	Signal segmentation	Part 1: 1D CNN and BiLSTM. Part 2: dense layer	NA	0.43
Varga *et al* (2018)	CHEST	Time & frequency. 9, 5 and 10 s length windows of EEG, EOG and AIRFLOW, respectively. Scaling SaO_2	2D convolution and two dense layers with 128 and seven outputs respectively	NA	0.42
Patane *et al* (2018)	CHEST, Chin1–Chin2, ECG, E1-M2, O2-M2, C4-M1, F4-M1 and O1-M2	Band pass filter on EEG, subsampling, normalization of overlapped segments and data augmentation	Six layers one-dimensional CNN architecture with 3/4 dense layers, Siamese architecture and of 3dense layers	Averaging of overlapped slices of segments and of AUPRC of 10-folds	0.40
Miller *et al* (2018)	NA	Signals -1-padded	Convolutional–deconvolutional neural network	NA	0.36
Warrick and Homsi (2018b)	NA	None	(ST- $LSTM_1$, and Dense.	Averaging per sample over 10 fold-classifiers	0.36
Present study	NA	None	ST, BN, DSC-1D, $BiLSTM_1$, BN, $BiLSTM_2$, BN and dense	Averaging per sample over 10 fold-classifiers	0.50

binary classes. All signals were padded with ones to a length slightly longer than the longest record. This approach achieved an AUPRC of 0.36.

2.2.2. Deep ensemble classifier

The top two approaches Howe-Þatterson *et al* (2018) and Práinsson *et al* (2018) employed an ensemble algorithm, suggesting that this was an important design decision for this problem.

Howe-Patterson *et al* (2018) used multiple dense convolutional units (DCU) and a BiLSTM layer. All PSG channels, excluding ECG, were anti-aliased by applying a finite impulse response (FIR) filter, down-sampled to 50 Hz and normalized over a 18 min moving window. Similar to Varga *et al* (2018), they did a simple scaling of the SaO_2 signal ensuring that physiologically relevant ranges were used. Four fold-classifiers were first constructed and then their corresponding predictions were averaged to estimate the final probability of sleep arousal regions. This ensemble yielded the top official-phase AUPRC of 0.54.

Þráinsson *et al* (2018) calculated time- and frequency-domain features of clinical and statistical origins. These were fed into a DNN with a BiLSTM, a dense layer with 50 outputs and a final dense-softmax output layer. They trained five classifiers using the same set of respiratory features, but with varying EEG-ECG channel selections. The final prediction was determined by averaging each classifier output per sample. This approach achieved an AUPRC of 0.45.

The approach presented in Patane *et al* (2018) consisted of four phases: preprocessing, windowing, data augmentation and classification. They preprocessed the EEG signals with a [0.5–45] Hz band-pass filter and removed candidate movement artefacts. All channels were segmented in 30 s time windows with 50% overlap and each segment was sub-sampled to 50 Hz and normalized. Data augmentation was also performed on-the-fly assuring that every batch of data had the same class proportion. Authors designed a deep architecture for multi-channel sleep arousal detection from EEG, EOG, EMG, AIRFLOW and SaO_2 signals (the ECG signal was excluded). It consisted of a six layer 1D-CNN architecture with three dense layers for processing most of the channels and with four dense layers for handling SaO_2 signal. The feature vectors were first merged together and fed through a Siamese architecture. Afterward, they used a sequence of three fully-connected layers to estimate the probability of sleep arousal for a whole window. The results for adjacent overlapping portions of segments were averaged together. Comparison of single- and multi-modal classifiers demonstrated that the multimodal model performing best while abdominal respiration (ABD) performed best in isolation. Their multimodal model AUPRC was 0.40.

Our 2018 Challenge base classifier consisted of two main components: a representation layer for feature extraction using the ST, and a sequencing learning layer composed of three LSTM layers with a dense-softmax layer to obtain classification predictions (Warrick and Homsi 2018b). An ensemble created by averaging the probabilities of ten base classifiers generated from cross-validation achieved an AUPRC of 0.36. Our challenge study also compared single and multi-modal classifiers with results in accordance with those of Patane *et al* (2018): the multi-modal classifier performed best, while the smaller set of features Chin-CHEST-ABD performed best in isolation. Our work and Miller *et al* (2018) were the only approaches that processed the entire signals as inputs without applying 'feature engineering', that is, without applying an explicit feature design and selection approach.

3. Methods

This section begins with an introduction to the 2018 Challenge dataset (Ghassemi *et al* 2018a, 2018b). A detailed description of the workflow to identify arousal regions in PSG signals is also described, which consists of three phases: data transformation, classification and evaluation.

3.1. Dataset

The development and evaluation of our proposed approach was based on the 2018 Challenge dataset. It consists of 1985 subjects monitored at the Massachusetts General Hospital (MGH) sleep laboratory for the diagnosis of sleep disorders. The data were partitioned into training ($n = 994$) and hidden test sets ($n = 989$). For each subject, 13 different physiological signals were collected during PSG sleep studies. The PSG recordings include six channels of EEG (F3-M2, F4-M1, C3-M2, C4-M1, O1-M2 and O2-M1), left eye EOG, an EMG lead placed at the chin (Chin1-2), respiratory movements placed on the chest and abdomen (CHEST and ABD), a single lead of ECG, SaO_2 and AIRFLOW. The sampling rate for all signals and labels was 200 Hz, sufficient to capture most of the PSG signal frequency bands mentioned in the introduction.

The training data was provided with five-class sleep-state labels and thirteen-class arousal-type labels. The sleep stages were annotated by clinical staff at the MGH according to the American Academy of Sleep Medicine manual for the scoring of sleep. The following six sleep stages were annotated in 30 s contiguous intervals: wakefulness, stage 1, stage 2, stage 3, rapid eye movement (REM), and undefined. Certified sleep technologists at the

MGH annotated waveforms for the presence of arousals that interrupted the sleep of the subjects. The annotated arousals were classified as either: spontaneous arousals, respiratory effort related arousals (RERA), bruxisms, hypoventilations, hypopneas, apneas (central, obstructive and mixed), vocalizations, snores, periodic leg movements, Cheyne–Stokes breathing or partial airway obstructions.

Target arousals were then defined as intervals from 2 s before a RERA arousal begins, up to 10 s after it ends or from 2 s before a non-RERA, non-apnea arousal begins, up to 2 s after it ends. Accordingly, the records for the training data were provided with three-class labels: arousal (A, +1), non-arousal (NA, 0) and unscored (NS, −1) regions.

3.2. Proposed solution

The main architecture of the base deep classifier shown in figure 1 consists of two main components; representation learning and sequencing learning, which are described below.

Our solution is designed in the spirit of full representation learning, i.e. learning directly from the raw data without any preprocessing. But because the training data corpus was 134 GB in size and we wanted a working solution on 'modest' computing resources (e.g. a system with 32 GB RAM and a single GPU with 12 GB of memory), we needed to make some tradeoffs. In the extreme, full representation learning from 200 Hz data is conceptually possible, but because the recordings are long (approximately 5M samples), training time would be onerous. Therefore our original motivation to use ST, a fixed nonlinear operation, was to provide a data reduction step before learning. LSTM training time is also a function of sequence length, so this first decimation step by a factor of 512 to a sampling rate of 0.5 Hz for the detection function was critical to a working solution.

3.3. Scattering transform

The scattering transform is a multidimensional, nonlinear function of real-valued signals (Bruna and Mallat 2013). In the past, this function has found practical applications in two different domains of signal processing: time series forecasting (Andreux 2018) and signal classification (Andén *et al* 2015). Interestingly, the task of arousal detection in PSG recordings lies at the interaction between these two perspectives. Indeed, sleep arousal can either be regarded, first, as a disruption in the predictability of the time series; or, alternatively, as a nonstationary pattern emerging from a stationary background. From this observation, we give hereafter two intuitive justifications of the interest behind adopting the scattering transform as a frontend to our deep learning system for sleep arousal detection. We refer to Lostanlen (2017) for further details in the argumentation.

First, in the context of time series forecasting, the scattering transform aims to encode, at every time step, past scalar values of a given sequence, and embed them into a *feature vector*, that is, another time series of fixed dimensionality. The purpose of this feature vector is to extract trends in the past values so that the prediction of the next few future values is relatively simple. In this respect, there is empirical evidence that the scattering transform reaches a favorable tradeoff between simplicity and accuracy: on a large class of real-world time series, the regression of scattering coefficients has a relatively high forecast accuracy, yet a relatively low intrinsic dimensionality (Andreux 2018).

In the specific case of the scattering representation, the feature space aggregates two subspaces, hereafter called 'orders'. The first order, which is comparable to a discrete wavelet transform (DWT), decomposes past trends into 11 typical time scales of periodicity which grow by factors of 2 from 10 ms to 10 s. As a result, if the stationary process underlying the observed time series were to consist of two independent linear trends of comparable periodicity, these would interfere within the same first-order subband. Resolving this interference is precisely the purpose of the second layer (Chudáček *et al* 2014). More precisely, second-order scattering decomposes, within each first-order subband, the temporal alternation between in-phase, constructive interference and out-of-phase, destructive interference, again into multiple time scales of periodicity. These second-order time scales are systematically greater than the 'baseband', first-order time scales. Therefore, the number of pairs of time scales that lead to a non-negligible outcome grows logarithmically with the second-order time scale. In other words, in consideration of the dimensionality of its feature vector, the scattering representation encodes its recent history with finer details, and its distant history with coarser details. This procedure of adaptation temporal resolution as a function of time lag is known in wavelet theory as foveation (Chang *et al* 2000).

Secondly, in the context of signal classification, the scattering transform performs a mapping of raw signal samples which reduces geometric, intra-class variability but retains informative, inter-class variability (Mallat 2016). In the case of PSG, the presence of small temporal lags and phase offsets between modalities is an example of such uninformative intra-class variability. Indeed, although PSG samples all modalities at a common rate of 200 Hz, spurious factors of experimental acquisition may cause a slight asynchrony between any two of the 13 channels. Likewise, swapping a pair of EEG electrodes leads to an inversion of electrical polarity, and thus a half-wave phase shift. None of these random modifications of data acquisition across different trials and subjects affect the task of interest. Therefore, it so appears that engineering a mid-level representation which, by design,

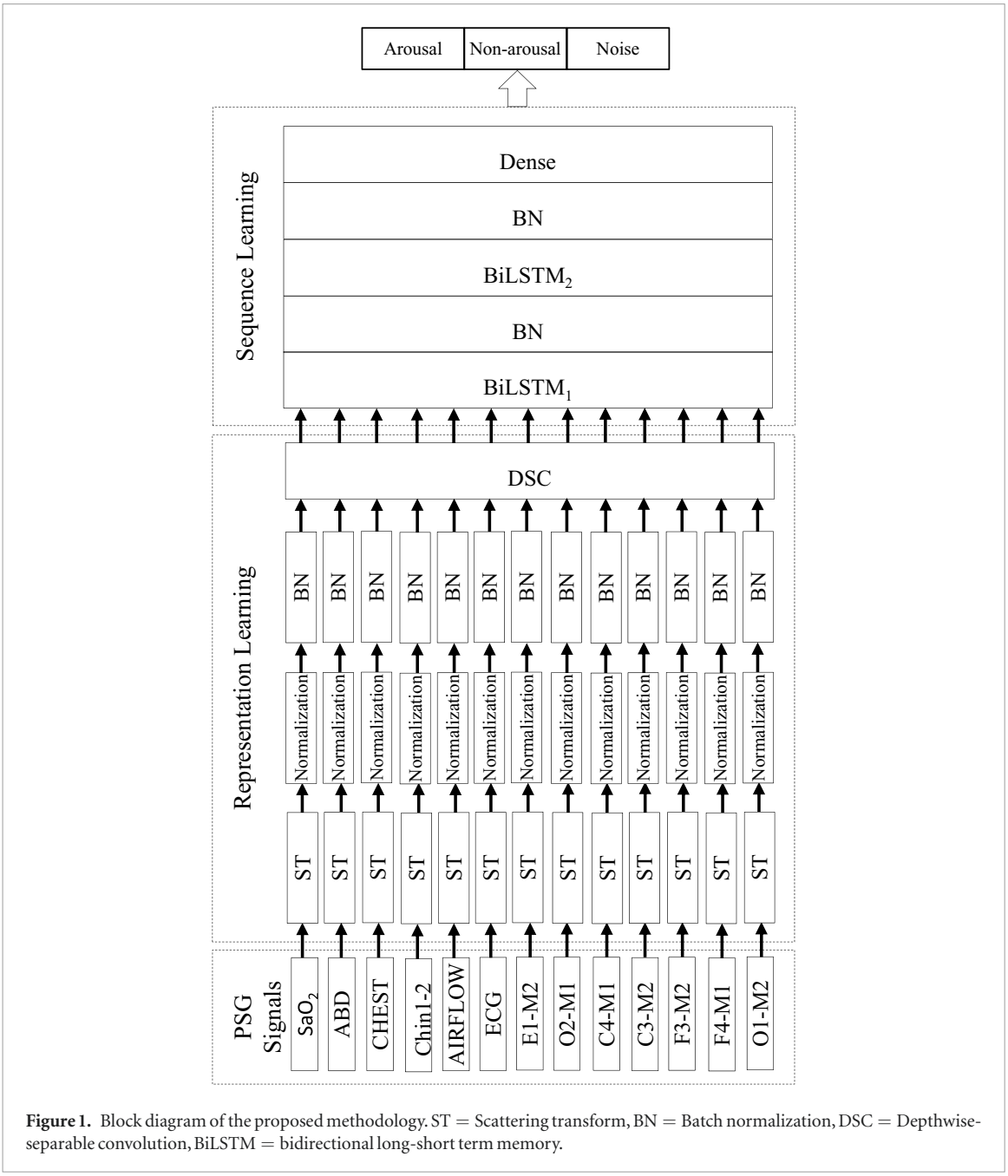

Figure 1. Block diagram of the proposed methodology. ST = Scattering transform, BN = Batch normalization, DSC = Depthwise-separable convolution, BiLSTM = bidirectional long-short term memory.

remains invariant to the action of channel-dependent phase shifts is beneficial to statistical robustness and generalization.

Given a time series of low-level, short-term features, the simplest way to ensure local invariance to time shifts is to apply a moving average. However, this local averaging procedure comes at the expense of losing fine details in highly oscillatory data. The rationale behind the scattering transform is, instead of building a linear invariant to translation in a single step, to perform multiple nonlinear operations of continuous wavelet transform and pointwise complex modulus, so as to demodulate the input signal into progressively slower modulation subbands, hereafter called scattering 'paths' (Mallat 2011). In the case of PSG, we apply two of such nonlinear wavelet modulus operators in a cascade before eventually convolving each path with a Gaussian window function. For all modalities, we set the time constant of this Gaussian window equal to 1 s. We used the open-source MATLAB library *scattering.m* (Lostanlen 2019) to compute the scattering coefficients.

In recent years, various authors in have proposed to train CNN models on scattering transform features for audio signal processing, thus forming hybrid ST-CNN pipelines (Peddinti *et al* 2014, Fousek *et al* 2015, Zeghidour *et al* 2016, Andreux and Mallat 2018). Conversely, there is a growing amount of literature on biomedical applications of the scattering transform, in conjunction with shallow learning methods (Chudáček *et al* 2014, Leonarduzzi *et al* 2018). Yet, prior to Warrick and Homsi (2018b), there had not been an empirical study on the applicability of deep learning from scattering coefficients of biomedical signals. Furthermore, there had not been any deep learning architecture, in any application domain, which trained long short-term memory networks

(LSTM) on scattering coefficients. We refer to Oyallon *et al* (2017) for a review of the state of the art in deep hybrid networks featuring both wavelet convolutions and trainable convolutions.

Another dimension in which our work is novel is the fact that we apply the scattering transform on multi-modal data. Although one prior study has applied scattering transforms to stereophonic audio (Lostanlen and Andén 2016), the question of training a machine learning model on the scattering representations of various sources of physiological data in simultaneity has remained considerably under-studied. In particular, the resort to depthwise-separable convolutions as an interface between multiple single-channel scattering coefficients with one multichannel LSTM network is a novel architectural contribution of ours.

3.4. Normalization

Because the generated ST coefficients had long exponential right tails, we performed a normalization step to transform them to quasi-normal distributions. This was done in two steps using the statistics of the training data for each ST path p of each channel e. For each coefficient $x[i, j, e, p]$ where i is the recording and j is the time sample, scaling by the median (calculated over all i and j) was followed by a log-like transformation $\mathcal{K}$ as in (1):

$$\text{Norm}(x[i, j, e, p]) = \mathcal{K}\Big\{\frac{\mu \cdot x[i, j, e, p]}{\text{median}(x[:, :, e, p])}\Big\}. \tag{1}$$

In our 2018 submission we had used $\mathcal{K}(x) = \log(x + 1)$. However, the Gaussian filter used to obtain the low frequency ST paths sometimes had transient negative values. We therefore used a modified transform $\mathcal{K}(x) = \sinh^{-1}(x) = \log[x + (x^2 + 1)]$ which admitted negative values. The constant μ was selected by searching for minimal skewness (i.e. quasi-normality). We first performed a search with coarse-grained resolution and then repeated the search with a finer resolution. Finally, we confirmed by visual inspection that the distributions were close to normal.

3.5. Classification phase

The arousal detection task from the 13 channel PSG recordings was carried out through two stages: building the base deep classifier and building the ensemble classifiers.

3.5.1. Depthwise-separable convolution

The ST output is hierarchically structured by channel $e \in \{1, ..., E\}$ and ST path $p \in \{1, ..., P\}$. In our 2018 Challenge entry we used $J = 8$ octave ST giving eight first-order and $\binom{8}{2} = 26$ second-order coefficients, such that $P = 36$. Our 2018 Challenge entry flattened the ST output as input to the subsequent LSTM layer, ignoring this structure.

Lower-frequency information, down to 0.1 Hz, is referred to in the EEG literature as the 'slow-frequency' band (Hiltunen *et al* 2014). In this new work we wished to capture this information which required $J = 11$. This increased P to $11 + \binom{11}{2} = 66$. This near two-fold increase in P would significantly augment the DNN model weight count. Furthermore, we wished to experiment with third-order ST which would add another $\binom{11}{3} = 165$ coefficients.

Therefore, we considered strategies to overcome the impact of the ST complexity on the model parameterization. We included this ST structure by factorizing the weights of the first LSTM layer into E electrodes $\times$ ST paths. Such an approach should also reduce rank, speed up training, and offer statistical regularization.

We chose a depthwise-separable convolution (DSC) (Chollet 2017) to address these issues. A DSC separates a standard convolution into two steps as shown in figure 2 for our configuration. We first perform the depthwise convolution X to mix the electrodes for each ST path with the $P \times E$ filter map F:

$$X[p] = \sum_e S[e, p]F[p, e]. \tag{2}$$

Then, the pointwise convolution mixes these transformed paths N times with $P \times N$ feature map G. N can be chosen arbitrarily and we choose $N = P$ to allow at least one output for each path p. Output Y results from applying the bias B and activation function ρ:

$$Y[n] = \rho\big(B[n] + \sum_p X[p]G[p, n]\big). \tag{3}$$

The total number of convolution coefficients including the bias weights is therefore $P \times E + (P + 1) \times N$. For $N = P = 66$ ST paths per channel, the DSC weight count is 5280. The main impact in overall weight count should occur in the subsequent BiLSTM layer. LSTM weight counts can be implementation-specific, so we tested with our Keras-Tensorflow subsystem for a precise comparison. We compared the weight counts of direct ST-BiLSTM layers to ST-DSC-BiLSTM layers using $N = P = 66$ and 100 cells per BiLSTM in each case. The former

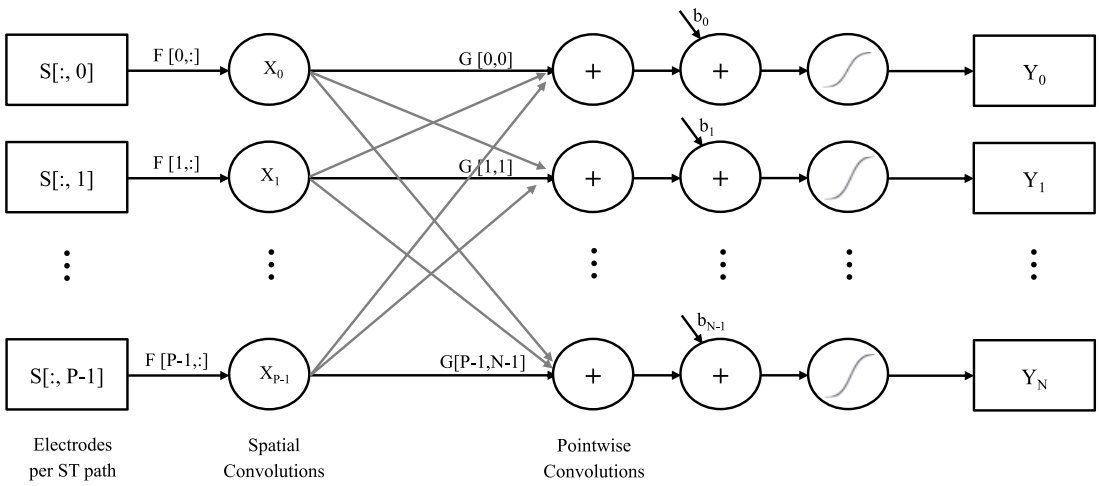

Figure 2. Depthwise-separable convolution of ST-transformed PSG signals.

had 456 000 weights in the BiLSTM layer while the latter had 134 400. Therefore, inserting the DSC layer between the ST and BiLSTM layers had the effect of reducing the weight count by 69% for this architecture.

3.5.2. *Sequence learning*

We chose LSTM units for arousal detection over 'memory-less' feedforward approaches because we hypothesized that the LSTM memory (short and long) might be important to interpret local and remote contexts. We had also used LSTMs in the 2017 Challenge (Warrick and Homsi 2018a) for ECG arrythmia detection where we demonstrated the value of both the long and short term memory of LSTM to that problem. We chose BiLSTM units over unidirectional LSTM because we anticipated that both past and future contexts of the PSG signals were important to detecting both arousal onset and termination. We had intended to use BiLSTMs for our challenge submission but were unable to do so because of time considerations: we noted that this was indeed a promising development direction in its successful use by the other 2018 Challenge groups.

We constructed models with the architecture of figure 1 using ten-fold cross-validation on the 2018 Challenge training data. The data was split into ten non-overlapping partitions to generate the test data for each fold. In each fold, we used $\frac{8}{9}$ of the test-complement data for training and $\frac{1}{9}$ for validation. The targets were decimated to the 0.5 Hz decision sampling rate and encoded as one-hot vectors for the three classes arousal, non-arousal and unscored. To prepare for efficient GPU batch training, both inputs and targets were end-padded to the length of the longest training recording.

With the above partitioning of folds, we had 793 training samples, 100 validation samples and either 99 or 100 test samples. We used RMSprop as the optimizer during training using the default Keras-Tensorflow parameters where the initial learning rate is set to 0.001 and is automatically adapted thereafter. We used a mini-batch size of 15 samples.

Then a classifier model was trained with early stopping: the best performing model (i.e. the one with the lowest validation loss) was retained in a checkpoint file. At the end of ten-fold cross-validation, we expected ten models having some degree of diversity due to the cross-validation permutations.

3.5.3. *Class imbalance*

Because arousals are relatively infrequent in the long sleep recordings, there is a considerable class imbalance that can present challenges to learning convergence, in terms of both iterations and accuracy. For this reason we weighted the loss function more heavily in arousal regions, experimenting with values close to the relative incidence of arousals (14) and with no emphasis (i.e. loss weight = 1). We fixed the loss weighting to 1 for all other classes.

3.5.4. *Auxiliary targets*

We hypothesized that the sleep-stage and arousal type labels marked by the experts might contain information useful to training in addition to the primary target (arousal state). We experimented using either the five-class sleep-stage labels or the thirteen-class arousal type labels as auxiliary targets during training. This was done by adding a dense-softmax layer parallel to that of the primary target after the BiLSTM layer. We used the sleep stage onsets and terminations to create a target vector at the decision sampling frequency. The arousal type stage target vector was created similarly. However, we found that these intervals overlapped in about 10% of the recordings; to resolve this conflict simply, we gave priority to intervals appearing later in time (overwriting when necessary).

The auxilary targets were encoded as one-hot vectors of length six and fourteen for sleep-stage and arousal type, respectively. The extra state represented undefined status in each case. For sleep stage we experimented with equal (1:1) and deemphasized (1:0.25) contributions to the overall loss function relative to the primary target. For arousal type we used equal contribution.

3.5.5. *Ensemble classifier*

Ensemble methods combine several trained classifiers into one predictive model with the objective to decrease variance (bagging), bias (boosting), or improve predictions (stacking) (Witten *et al* 2016, Ju *et al* 2018). In our approach, we built our ensemble classifier by using the unweighted averaging strategy to fuse the decision of ten base deep classifiers. This rule chooses the class with the highest average probability over all models $\mathcal{M}_i$, $i \in \{1, \ldots, n\}$, where n in this context is the number of cross-validation folds.

3.6. Evaluation

The proposed classifier was evaluated with two metrics that give complementary insights into the classification performance: AUPRC and area under the receiver operating characteristic (AUROC). We relied on the Python code provided by the 2018 Challenge to compute these metrics, which calculates histograms to estimate scaled versions of the two conditional prediction probabilities $Prob[arousal|(truth = \text{arousal})]$ and $Prob[arousal|(truth = \text{non-arousal})]$. These histograms are used to determine the false-positive rates, recall and precision values for each probability threshold, taken from corresponding histogram bins; we used the default setting of 1000 histogram bins. These allow us to generate the ROC and PR curves and estimate their areas.

For performance reasons, we compared targets and predictions at the 0.5 Hz decision sampling rate in all our experiments. For final evaluation on the Challenge server, we up-sampled the predictions by sample and hold to the expected 200 Hz rate.

As a baseline comparison of ST processing with conventional spectral methods, we also performed an experiment using power spectral density (PSD) analysis. This was done by first detrending the 200 Hz signals with a FIR high-pass filter having a cutoff frequency of 1 Hz. Then we performed sliding-window analysis with a window size of 512 samples that was advanced by the same amount, thus conforming to the decision sampling rate of the ST analysis. For each window we created a model using the matlab function pcov using an FFT length of 64, generating approximately the same number of coefficients (33) as the ST36 processing, A fixed AR model order of 20 was chosen empirically. The frequency resolution provided by the PSD coefficients was therefore 200 Hz / 64 = 3.1 Hz. For the very slow-changing SaO_2 signal, the above processing was inappropriate, and the coefficient was set to the window average. We performed the normalization as before using the $\log(x + 1)$ transform. Finally, we trained a detector in cross-validation using a three-layer BiLSTM for fair comparison.

Differences between experiment results were often subtle (within a few percent mean AUPRC and AUROC), so we needed a principled way to compare two experiments. We did this by performing a 2-sided t-test of test AUPRC over all ten folds to test the null hypothesis that the mean AUPRCs were equal.

4. Results

4.1. Scattering

Figure 3(a) shows the normalized ST coefficients for all channels for a typical recording. Figure 3(b) is a magnified view that includes a region of arousals. It is apparent from these overall and magnified views that the onset and termination of the arousals appear as higher ST values (i.e. towards red on the color maps). It is also clear that there are numerous elevated values where there are no target arousals. These clearly appear itermittently in non-scored (noisy) regions (e.g. between sample 3.4×10^6 and 4.4×10^6) but they may also indicate sensitivity to non-target arousals. Finally it is apparent from figure 3(b) that the time support of the second-order coefficients (near the bottom of each channel map) is wider than those of the first-order coefficients (at the top of the map), and especially with decreasing frequency band (lower in the map).

4.2. Depthwise-separable convolution

Figures 4 and 5 show the DSC depthwise and pointwise filter-weight maps, respectively, for an example fold of the cross-validation. They are presented as 11-by-11 lower-triangular images with first-order paths along the diagonal and second-order paths beneath; table 2 gives the Morlet filter octave(s) associated with each path in the matrix.

For the depthwise convolution, all channels have non-zero weights for at least some paths; however, the four channels ABD, CHEST, AIRFLOW and SaO_2 seem to dominate over the seven EEG channels, Chin1-2 and ECG: their weights tends to more towards the red (positive) and blue (negative) extremes of the colormap. While these four channels also had non-zero weights in the higher octaves (towards the upper left), their lower octave weights (toward the lower right) were especially strong.

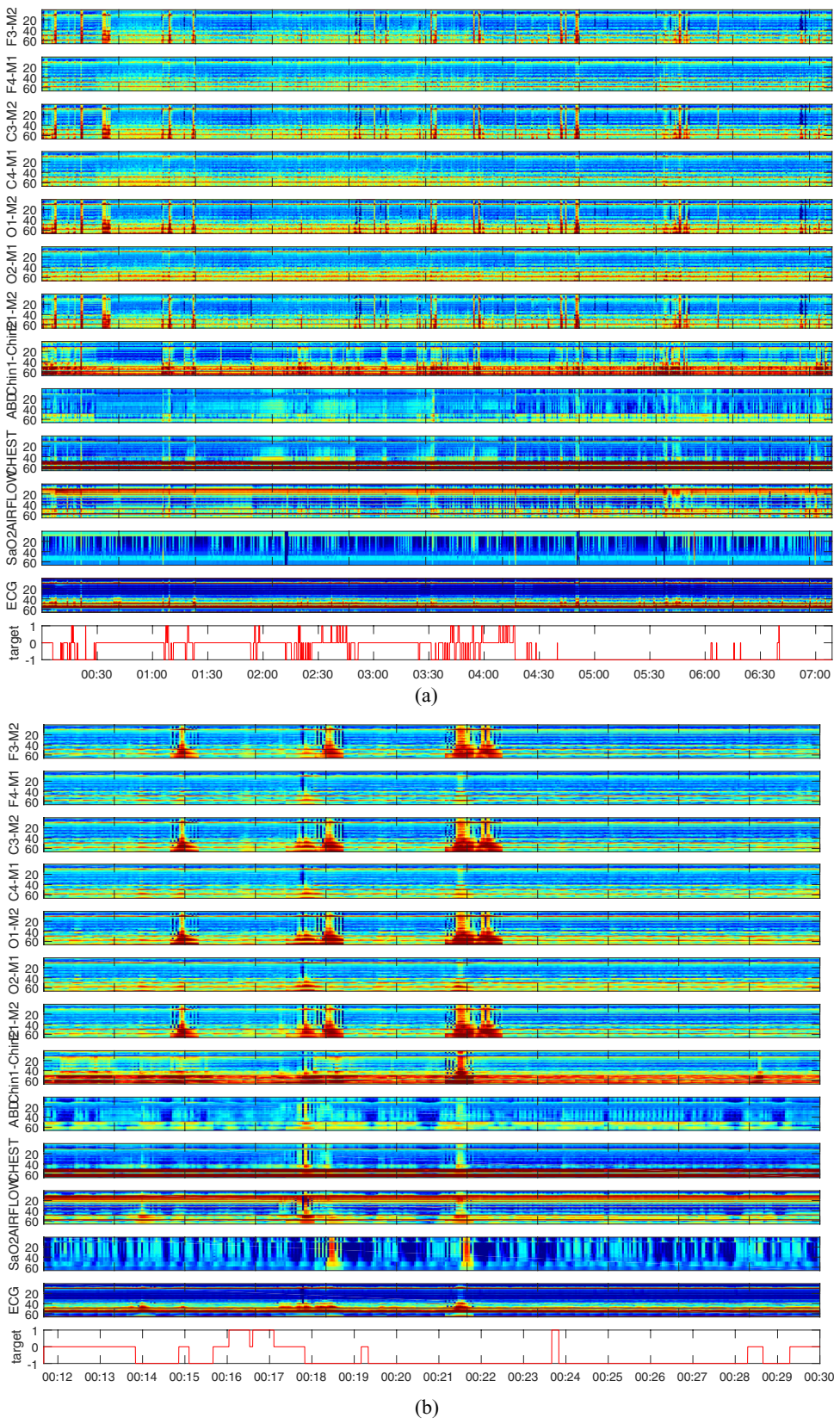

Figure 3. (a) Typical scattering transforms and target for entire recording tr03-0005. For each channel, there are 11 1st-order and 55 2nd-order coefficients (concatenated vertically). Normalized values range from blue (lowest) to red (highest). Target values of −1, 0 and 1 indicate not-scored, non-arousals and arousals, respectively. (b) Magnified from the first 30 min of (a). The time axes are shown in hour:minute format. Clusters of high ST values are apparent in the vicinity of target arousals.

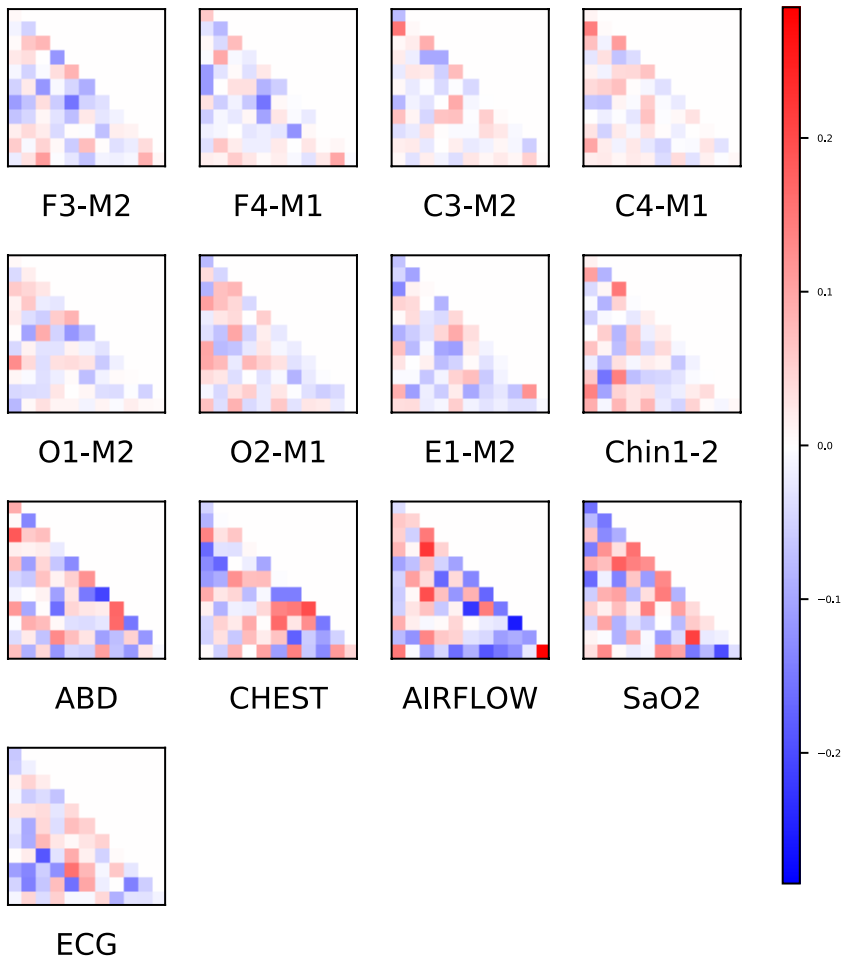

Figure 4. Depthwise filter F for fold 1 with octave-pair ST path maps for each channel. The paths are arranged as in table 2.

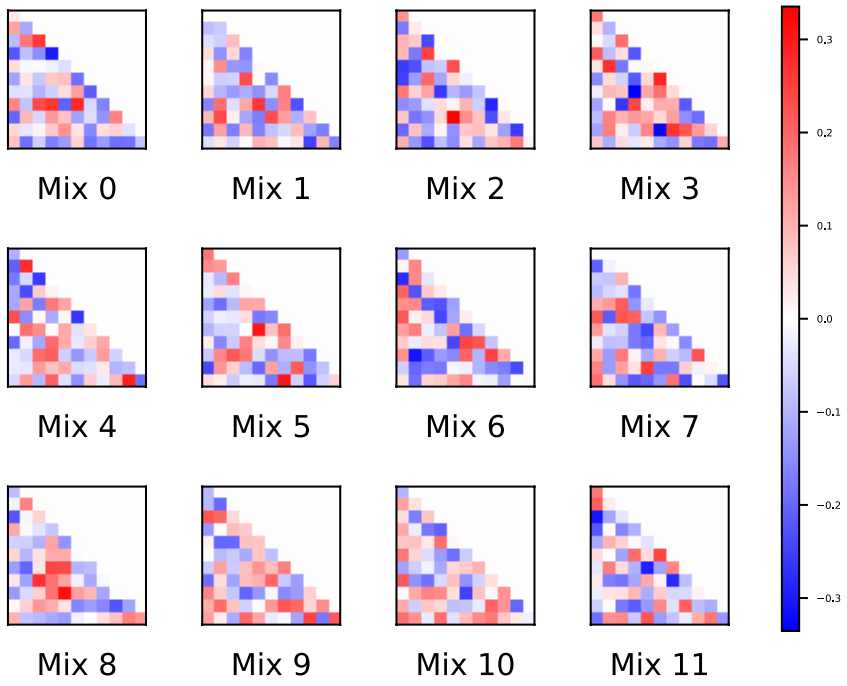

Figure 5. Pointwise filter G of fold 1 for the first 12 of the $N = 66$ mixtures of first-order and second-order (octave-pair) ST paths. The paths are arranged as in table 2. Note that the colormap scale here differs from figure 4.

Table 2. Position of each scattering path $i \in \{0, 1, \ldots, (P-1) = 65\}$ in a matrix of Morlet filter octaves (in bold italic). The first order paths appear along the diagonal (in italic) while the second-order paths appear below the diagonal (in bold).

100	0								*First order*		
50	**11**	*1*							**Secord order**		
25	**12**	**13**	*2*								
12.5	**14**	**15**	**16**	*3*							
6.25	**17**	**18**	**19**	**20**	*4*						
3.125	**21**	**22**	**23**	**24**	**25**	*5*					
1.56	**26**	**27**	**28**	**29**	**30**	**31**	*6*				
0.781	**32**	**33**	**34**	**35**	**36**	**37**	**38**	*7*			
0.390	**39**	**40**	**41**	**42**	**43**	**44**	**45**	**46**	*8*		
0.195	**47**	**48**	**49**	**50**	**51**	**52**	**53**	**54**	**55**	*9*	
0.097	**56**	**57**	**58**	**59**	**60**	**61**	**62**	**63**	**64**	**65**	*10*
Octave (Hz)	***100***	***50***	***25***	***12.5***	***6.25***	***3.125***	***1.56***	***0.781***	***0.390***	***0.195***	***0.097***

AIRFLOW, for example had strong first-order 0.097 Hz (intense red) and 0.39 Hz (intense blue) weights as well as strong second-order weights for the 12.5/25 Hz (red) and 1.56/25 Hz (red) and 0.781/3.125 Hz (blue) octave pairs. SaO_2 had higher-octave contributions in addition to its strong lower octave contributions, possibly reflecting sensitivity to sensor disturbance.

Weights were diversely spread across the seven EEG channels. In particular, lower octaves had significant weights over several electrodes: the first-order 0.197 Hz octave was present in E1-M2, and the second order 0.097/0.197 octave pair was present for F3-M2 and F4-M1 and to a lesser degree, C3-M2. ECG was especially strong for the 0.39/6.25 Hz (red) and 0.781/25 Hz (blue) octave pairs. For the Chin1-2 EMG signal, the first-order 25 Hz (red) and second order 0.39/50 Hz octave pair (blue) stood out.

For brevity we show only 12 of the $P = 66$ mixtures of the pointwise filter in figure 5. These maps are more difficult to interpret, with strong peaks (red) and valleys (blue) occuring in many of the first and second order paths, and with much diversity between mixes.

4.3. Training and classification

Table 4 shows the results of our classifier training experimentation. The models were compared incrementally to assess the impact of the introduction of each change. We used a slightly modified version of our 2018 Challenge entry for ST36-UniLSTM (Exp. 1). It consisted of second-order ST cofficients spanning $J = 8$ octaves ($P = 36$ ST paths per channel), three unidirectional LSTM layers, each followed by a BN layer, and a dense-softmax layer. Unlike the Challenge entry, however, we used all three arousal targets {A, NA, NS} rather than just two {A, NA} to allow more noise-related expressiveness in the model. The mean number of training epochs at convergence was 140, with training time on the order of 16 hrs per fold. The mean $\pm$ standard deviation of the fold AUPRC was 0.2913 ± 0.0398.

Next, in Exp. 2 we replaced the LSTM units with BiLSTM. This time the AUPRC was 0.3753 ± 0.0398, an increase of 28.8% over baseline ($p = 0.0002$). Training time was similar to baseline, but the mean epochs at minimum loss reduced to 112.6.

The baseline PSD-BiLSTM detector of Exp. 3 was inferior to the equivalent ST36-BiLSTM detector of Exp. 2, with AUPRC decreasing from 0.3753 to 0.2839 ($p < 0.0001$).

We then augmented the ST analysis to capture the 'slow-frequency' band (0.1 Hz), increasing P to 66 (Exp. 4). The AUPRC was 0.4074 ± 0.0502, an increase of 8.5% over the Exp. 2 ($p = 0.0141$). The rate of training convergence was similar.

Then a DSC layer with $N = P = 66$ and linear activation was inserted between the ST and first BiLSTM layers. The change in AUPRC was insignificant (Exp. 5, $p = 0.9078$); however, when the DSC activation was changed to a ReLU unit and a BN layer was added to the input, the AUPRC increased to 0.4344 ± 0.0369, an increase of 6.6% (Exp. 6, $p = 0.0395$). As well, addition of the DSC reduced training time considerably in both cases; in the latter case, it reduced to a mean of 24 epochs at minimum loss and 10 hrs per fold. The rationale for adding the initial BN layer was that while normalization successfully transformed the ST coefficients to a quasi-normal distribution, the mean was considerably offset (approximately in the range of 6–10). Incorporating mean removal into the normalization step would have been the ideal approach, but to avoid renormalizing the data before training, we used the BN layer to accomplish the same task during training.

Removing the increased loss weighting on arousal regions by setting this value to 1 increased the AUPRC to 0.4547 ± 0.0308, an increase of 4.7% (Exp. 7, $p = 0.0391$). Convergence was slightly longer, however, with mean epochs at minimum loss of 31.

As a further study in model ablation, we reduced the number of BiLSTM layers from three to two. The AUPRC did not change significantly (Exp. 8, $p = 0.8121$), but convergence was slower, with the mean epoch at minimum loss increasing slightly to 42. We retained this simpler model in subsequent experiments.

Finally, we added the auxiliary target of arousal type to the network architecture, giving it equal weight in the loss function to the primary arousal target. This increased AUPRC to 0.4675 ± 0.0444, an increase of 2.4% (Exp. 9, $p = 0.0741$). The mean epoch at minimum loss increased to 50, likely due mostly to the additional calculation of gradients for this 15-class, one-hot auxiliary target. Training and validation AUPRC were 0.6092 ± 0.0246 and 0.4608 ± 0.0175, respectively. The test, training and validation AUROCs were 0.9180 ± 0.0066, 0.9500 ± 0.0047 and 0.9230 ± 0.0050, respectively.

We considered this to be our best architecture. We used the models generated from 10-fold cross-validation in our proposed ensemble. In a 2018 Challenge follow-up entry, this ensemble scored 0.50 on the hidden test set.

Other experiments that did not significantly change performance included adding third-order ST coefficients, inserting 20% dropout layers before or after the DSC layer and using the sleep stages as auxiliary targets.

Figure 6 shows sample receiver operating characteristic (ROC) curves and precision-recall (PR) curves using a single fold of our final architecture. Since our final classifier model had relatively small variance (standard deviation of 0.0444) across the ten folds of cross-validation, as shown in table 4, we considered the ROC and PR curves of one of these folds (fold 1) to be representative; we chose a decision probability threshold ($p = 0.17$) based on an example criteria of maximum F1 (0.48) using the validation data set. F1 is the harmonic mean of precision and recall, defined as $2 \cdot \frac{precision \cdot recall}{precision + recall}$. Applying this threshold to the test data resulted in a false-positive rate of 0.06, a recall of 0.61 and a precision of 0.44. Table 3 tabulates these performance measures on the validation data as a function of probability threshold at 0.05 recall intervals.

5. Discussion

The results show an overall improvement of mean AUPRC in cross-validation from 0.2913 to 0.4675 compared to our 2018 Challenge entry. The ensemble amplified this improvement to 0.50 when used on the independent data of the hidden test set.

The most important contributions to increased performance were, in decreasing impact: the use of BiLSTM; the addition of the slow-frequency ST coefficients; the combination of Input BN, DSC and ReLU; uniform weighting of loss function for all targets; and addition of auxiliary target arousal type.

The importance of the BiLSTM indicates that future context is critical to the arousal detection, and likely to the accurate delineation of arousal termination. The use of a BiLSTM is well suited to PSG interpretation because it is typically satisfactory to provide offline processing, after a night sleep study is complete, for example. BiLSTM is less suitable in realtime, low-latency contexts.

In our 2018 Challenge entry we used eight-octave ST coefficients to capture frequencies as low as 0.75 Hz. We have demonstrated in this study that frequencies in the lower 0.1 Hz slow-frequency band are discriminative for arousal detection. Definitions of the lowest (delta) band in classical EEG investigation vary, with some defining it as 0.5 Hz-4 Hz and others 1.0 Hz–4 Hz. But there is certainly EEG activity below that, labelled the 'slow-frequency' (0.1–1 Hz) and 'infra-slow' bands (0.01–0.1 Hz). These bands have been found to be correlated to 'resting states' in studies with simultaneous acquisition of EEG and MRI (Hiltunen *et al* 2014). This is an interesting finding that may support the slow-frequency band discrimination that we have observed. In a future study, it would be interesting to explore the order-of-magnitude lower infra-slow band. The ST-DSC-BiLSTM approach that we describe could be adapted to this purpose in a straightforward way.

This study demonstrates that the ST is a powerful fixed-base transform for the multimodal biomedical signals of PSG. Although the PSD detector was not tuned for performance, it was a helpful baseline implementation. The ST36-BiLSTM detector, having an equivalent number of ST coefficients, clearly outperformed its PSD33-BiLSTM equivalent. This is consistent with studies in Bruna and Mallat (2013) which indicated that for equivalent image classifiers, first-order ST coefficients alone outperformed windowed Fourier transform coefficients as inputs, and addition of second-order ST consistently performed even better over a range of noise level conditions.

In addition to contributing to improved performance, the DSC offered the advantages of greater training efficiency and explicit consideration of the structure in the data and the reduction of correlations related to this structure. This makes sense as nearby electrodes of the same modality (especially the eight channels of EEG) are highly correlated; this is addressed by the depthwise convolution. It also seems reasonable to expect that the pointwise convolution reduces correlations between ST paths in close frequency proximity within the same modality as well as between modalities.

The DSC weights thus provide key insights into which electrodes and paths are most informative (and therefore discriminative). Both first- and second-order ST paths were important in the DSC filters, as demonstrated

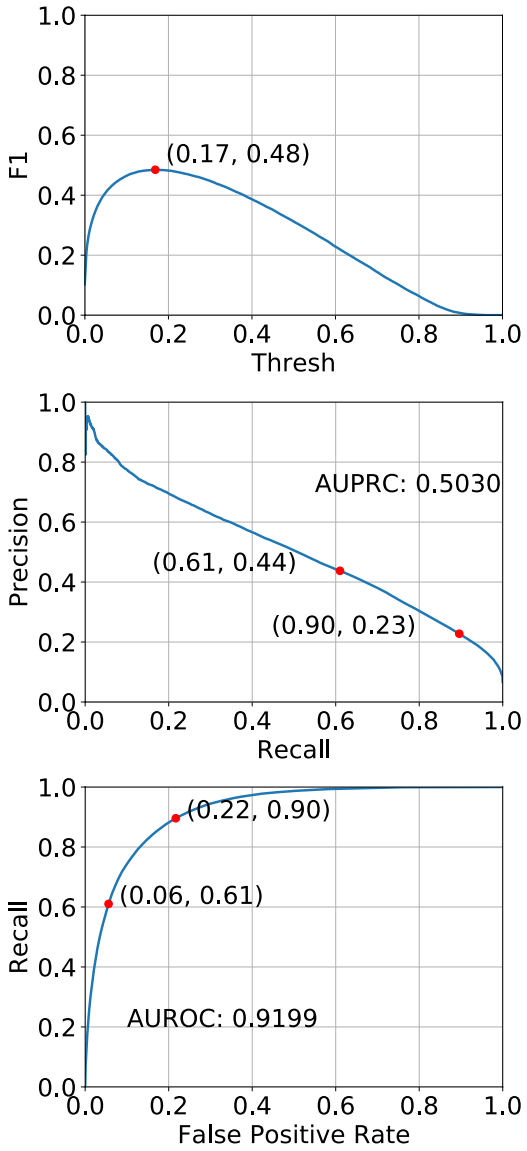

Figure 6. Performance based on threshold selection for fold 1 (top) Validation data F1 versus prob. threshold: the selected threshold of $p = 0.17$ occurs at maximum F1. (middle) Test data PRC (bottom) Test data ROC. The points in red indicate performance at selected thresholds.

both in the depthwise mixtures of electrodes over the same path (figure 4) and the pointwise mixtures of path information (figure 5)

The depthwise filter weights for the four-channel set ABD, CHEST, AIRFLOW and SaO_2 seem to dominate the other channels and therefore were important for arousal discrimination. Their lower-octave contributions were especially strong, consistent with Nino *et al* (2013) who found evidence of apnea arousal activity in these four channels in the 0.05 and 0.2 Hz frequency ranges. These phenomena may also be relevant to the non-apnea arousals that concern this study.

The EEG channels weights were less strong (i.e. they had fainter colors in the filter maps), but they were diversely spread across these seven channels, likely reflecting their varying spatial sensitivities to the energies of the dipole sources at each octave. It is noteworthy that while the addition of the three octaves down to as low as the slow-frequency band had the greatest impact on the four channel set ABD, CHEST, AIRFLOW and SaO_2, it did also increase the EEG information for these octaves (confirming the original motivation for their addition). We also note that the ECG peak response at 0.781 Hz closely corresponds the the spectral peak of the sinus rhythm at approximately 1.0 Hz (Clifford *et al* 2006). Finally, while higher EMG frequencies were dominant, informative activity below the limit of 5 Hz reported in Viitasalo and Komi (1977) was also apparent.

What is striking about the pointwise filter maps of figure 5 is less about their interpretability and more about their diversity, demonstrating their flexibility to learn discriminating clusters of coexisting paths. Further hyper-parameter tuning is required to determine whether the $N = 66$ paths are sufficient or excessive to adequately present the transformed input signals to the subsequent sequence-learning layers.

Table 3. Table of performance measures versus probability threshold for the validation data. FP Rate refers to the false positive rate and Prec refers to the precision. Two proposed thresholds are shown in bold.

p	Recall	FP Rate	Prec	F1
0.00	1.00	1.00	0.05	0.10
0.01	0.95	0.27	0.17	0.28
0.02	**0.90**	**0.19**	**0.21**	**0.35**
0.04	0.85	0.14	0.26	0.39
0.06	0.80	0.11	0.29	0.42
0.08	0.75	0.09	0.32	0.45
0.10	0.70	0.07	0.35	0.47
0.13	0.65	0.06	0.38	0.48
0.16	0.60	0.05	0.41	0.48
0.17	**0.58**	**0.05**	**0.42**	**0.48**
0.22	0.50	0.03	0.46	0.48
0.25	0.45	0.03	0.49	0.47
0.29	0.40	0.02	0.52	0.45
0.34	0.35	0.02	0.55	0.43
0.39	0.30	0.01	0.58	0.39
0.44	0.25	0.01	0.62	0.36
0.51	0.20	0.01	0.65	0.31
0.58	0.15	0.00	0.70	0.25
0.66	0.10	0.00	0.74	0.18
0.76	0.05	0.00	0.80	0.09

Table 4. Relative impact of cumulative architectural or training changes in cross-validation experiments. Mean AUPRC comparisons are made with the previous experiment except where indicated with asterisks: Exp. 4 was compared to 2 and Exp. 6 was compared to 4. Exp. 9 (in bold) was our best model submitted for hidden-test results in the followup phase.

		Test AUPRC				Mean fold training	
Exp.	Change	mean	Std	% change	p-value	Epochs	∼Hrs
1	ST36-UniLSTM	0.2913	0.0398	—	—	140.2	16
2	ST36-BiLSTM	0.3753	0.0398	+28.8	0.0002	112.6	16
3	PSD33-BiLSTM	0.2839	0.0187	−24.4	<0.0001	26.0	6
4	ST66('slow freq.')	0.4074	0.0502	+8.5∗	0.0141	104.9	16
5	DSC-linear	0.4089	0.0289	+0.4	0.9078	29.8	10
6	Input BN-DSC-ReLU	0.4344	0.0369	+6.6∗	0.0395	23.9	10
7	No arousal loss weight	0.4547	0.0308	+4.7	0.0391	30.8	10
8	$(\text{BiLSTM-BN})^2$	0.4564	0.0359	+0.4	0.8121	42.1	10
9	**Auxiliary target arousal type**	**0.4675**	**0.0444**	**+ 2.4**	**0.0741**	**50.2**	**10**

The addition of the auxiliary arousal type (Exp. 7 in table 4) increased the mean test AUPRC by 2.4% compared to a model without it (Exp. 6) although it did so with less significance ($p = 0.0741$) than the other influential changes mentioned above. The addition of these 13 arousal types may have influenced training in such a way as to correctly distinguish some borderline arousal regions from background. It is interesting to note that other more 'neutral' changes with respect to mean AUPRC such as Exp. 4 and 7 had much higher p values (0.9078 and 0.8121, respectively). Given that this architecture produced the highest mean test AUPRC, we used it in our final followup phase submission.

In the early stages of our work, our training infrastucture was insufficient to train on all the data. For that reason we performed smaller batches of training, for example with 200 recordings. Under these conditions, we found that augmenting the arousal loss weight was essential to successful training. We showed in this work, where training of the entire training set was possible, that our final model performed better without such loss emphasis. This was an interesting result because it demonstrated that given enough data, the ST-DSC-BiLSTM approach could train very successfully with the data 'as-is', without resorting to heuristics such as loss weighting, subsampling or augmentation to counteract class imbalance. Being able to process long time series continuously in this way is critical to capturing long-term trends or sentinel events that are temporally remote. Our success with PSG signals gives some confidence that this approach could perform well without ad hoc adjustments in other time-series contexts.

To be clinically useful, it is necessary to choose some decision threshold to produce a classifier instance, rather than the family of classifiers described by AUROC and AUPROC. The curves of figure 6 illustrate one representative selection. The maximum F1 criterion for threshold selection considers recall (sensitivity) and precision (PPV) to have equal value. But other criteria are possible, notably if, in our context of arousal detection, there is tolerance for higher false positive rates to achieve greater sensitivity: if the system is used to screen long records for a human to edit, for example. As a plausible threshold selection for this scenario, table 3 and figure 6 indicate that to achieve a recall of 0.90 requires a threshold of $p = 0.02$; under these conditions an estimated false positive rate of 0.19 and a precision of 0.21 would result.

6. Conclusions

This study confirms that the proposed representation learning layer, which consists of scattering transforms of each channel and depthwise-separable convolution as an interface to the LSTM-based sequential learning, had a substantial impact on the performance of the pipeline for the detection of arousal regions in PSG recordings. Experiments showed that capturing lower-frequency information, down to 0.1 Hz, and using a stack of BiLSTM layers yielded a AUPRC of 0.50, a substantial increase of 0.14 over our previous approach submitted during the official phase. Future work will be directed at applying this approach to the infra-slow band (down to 0.01 Hz), an order of magnitude lower in frequency than the work of this paper yet a straightforward extension of our approach. This has the potential to improve the low-frequency representation for all the (non-EMG) PSG signals and achieve increased discrimination.

In the follow-up phase of the 2018 PhysioNet/CinC Challenge the proposed architecture achieved a state-of-the-art AUPRC of 0.50 on the hidden test data, tied for the second-highest official result overall.

Acknowledgment

This work is partially supported by the ERC InvariantClass Grant No. 320959. The authors would like to acknowledge the computational facilities provided by PeriGen Inc. for this work.

ORCID iDs

Philip A. Warrick https://orcid.org/0000-0002-6945-6271
Vincent Lostanlen https://orcid.org/0000-0003-0580-1651
Masun Nabhan Homsi https://orcid.org/0000-0001-7427-6198

References

AASM 1992 EEG arousals: scoring rules and examples *Sleep* **15** 174–84
Agarwal R 2006 Automatic detection of micro-arousals *IEEE Engineering in Medicine and Biology 27th Annual Conf.* (IEEE) pp 1158–61
Aggarwal K, Khadanga S, Joty S, Kazaglis L and Srivastava J 2018 A structured learning approach with neural conditional random fields for sleep staging *IEEE Int. Conf. on Big Data* (IEEE) pp 1318–27
Andén J, Lostanlen V and Mallat S 2015 Joint time-frequency scattering for audio classification *Proc. MLSP*
Andreux M 2018 Modélisation autorégressive, fovéale et neuronale de séries temporelles *PhD Thesis* École normale supérieure (in French)
Andreux M and Mallat S 2018 Music generation and transformation with moment-matching scattering inverse networks *Proc. ISMIR*
Bruna J and Mallat S 2013 Invariant scattering convolution networks *IEEE Trans. Pattern Anal. Mach. Intell.* **35** 1872–86
Chang E C, Mallat S and Yap C 2000 Wavelet foveation *Appl. Comput. Harmon. Anal.* **9** 312–35
Cho S P, Lee J, Park H and Lee K 2006 Detection of arousals in patients with respiratory sleep disorders using a single channel EEG *IEEE Engineering in Medicine and Biology 27th Annual Conf.* (IEEE) pp 2733–5
Chollet F 2017 Xception: deep learning with depthwise separable convolutions *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* pp 1251–8
Chudáček V, Andén J, Mallat S, Abry P and Doret M 2014 Scattering transform for intrapartum fetal heart rate variability fractal analysis: a case-control study *IEEE Trans. Biomed. Eng.* **61** 1100–8
Clifford G, Azuaje F and McSharry P 2006 *Advanced Methods and Tools for ECG Data Analysis* (*Artech House Engineering in Medicine & Biology Series*) (Boston, MA: Artech House)
De Carli F, Nobili L, Gelcich P and Ferrillo F 1999 A method for the automatic detection of arousals during sleep *Sleep* **22** 561–72
Fousek P, Dognin P and Goel V 2015 Evaluating deep scattering spectra with deep neural networks on large scale spontaneous speech task *Proc. IEEE ICASSP* pp 4550–4
Ghassemi M M, Moody B E, Li-Wei H L, Song C, Li Q, Sun H, Mark R G, Westover M B and Clifford G D 2018a You snooze, you win: the PhysioNet/Computing in Cardiology Challenge 2018 *Computing in Cardiology* vol **45**
Ghassemi M M, Moody B E, Li-Wei H L, Song C, Li Q, Sun H, Mark R G, Westover M B and Clifford G D 2018b You snooze, you win: the PhysioNet/Computing in Cardiology Challenge 2018 (http://physionet.org/challenge/2018) (Accessed: May 11, 2019)
Halász P, Terzano M, Parrino L and Bódizs R 2004 The nature of arousal in sleep *J. Sleep Res.* **13** 1–23
He R, Wang K, Zhao N, Liu Y, Yuan Y, Li Q and Zhang H 2018 Identification of arousals with deep neural networks (DNNs) using different physiological signals *Computing in Cardiology* vol **45**
Hiltunen T *et al* 2014 Infra-slow EEG fluctuations are correlated with resting-state network dynamics in fMRI *J. Neurosci.* **34** 356–62

Howe-Patterson M, Pourbabaee B and Benard F 2018 Automated detection of sleep arousals from polysomnography data using a dense convolutional neural network *Computing in Cardiology* vol **45**

Huupponen E, Væ rri A, Hasan J, Saarinen J and Kaski K 1996 Sleep arousal detection with neural network *Proc. 1st Int. Conf. on Bioelectromagnetism Medical & Biological Engineering & Computing* pp 219–20

Jarvis M R and Mitra P P 2000 Apnea patients characterized by 0.02 Hz peak in the multitaper spectrogram of electrocardiogram signals *Computers in Cardiology* vol **27** (Cat. 00CH37163) pp 769–72

Jégou S, Drozdzal M, Vazquez D, Romero A and Bengio Y 2017 The one hundred layers tiramisu: fully convolutional densenets for semantic segmentation *Proc. IEEE Conf. on Computer Vision and Pattern Recognition Workshops* pp 11–9

Ju C, Bibaut A and van der Laan M 2018 The relative performance of ensemble methods with deep convolutional neural networks for image classification *J. Appl. Stat.* **45** 1–19

Leonarduzzi R, Abry P, Wendt H, Kiyono K, Yamamoto Y, Watanabe E and Hayano J 2018 Scattering transform of heart rate variability for the prediction of ischemic stroke in patients with atrial fibrillation *Methods Inf. Med.* **57** 141–5

Lostanlen V 2017 Convolutional operators in the time-frequency domain *PhD Thesis* École normale supérieure

Lostanlen V 2019 Scattering.m a matlab toolbox for signal scattering (https://github.com/lostanlen/scattering.m) (Accessed: May 11, 2019)

Lostanlen V and Andén J 2016 Binaural scene classification with wavelet scattering *Proc. DCASE*

Mallat S 2011 Group invariant scattering *Commun. Pure Appl. Math.* **65** 1331–98

Mallat S 2016 Understanding deep convolutional networks *Phil. Trans. R. Soc.* A **374** 20150203

Miller D, Ward A and Bambos N 2018 Automatic sleep arousal identification from physiological waveforms using deep learning *Computing in Cardiology* vol **45**

Niedermeyer E and da Silva F 2004 *Electroencephalography: Basic Principles, Clinical Applications, and Related Fields* (Philadelphia, PA: Lippincott Williams & Wilkins)

Nino C L, Rodriguez-Martinez C E, Gutierrez M J, Singareddi R and Nino G 2013 Robust spectral analysis of thoraco-abdominal motion and oxymetry in obstructive sleep apnea *Conf. Proc.: ... Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conf.* vol 2013 pp 2906–10

Oyallon E, Belilovsky E and Zagoruyko S 2017 Scaling the scattering transform: Deep hybrid networks *Proc. IEEE Int. Conf. on Computer Vision* pp 5618–27

Patane A, Ghiasi S, Scilingo E and Kwiatkowska M 2018 Automated recognition of sleep arousal using multimodal and personalized deep ensembles of neural networks *Computing in Cardiology* vol **45**

Peddinti V, Sainath T, Maymon S, Ramabhadran B, Nahamoo D and Goel V 2014 Deep scattering spectrum with deep neural networks *Proc. IEEE ICASSP* pp 210–4

Þráinsson H, Ragnarsdóttir H, Kristjánsson G and Marinósson B 2018 Automatic detection of target regions of respiratory effort-related arousals using recurrent neural networks *Computing in Cardiology* vol **45**

Saeed A, Trajanovski S, Van Keulen M and Van Erp J 2017 Deep physiological arousal detection in a driving simulator using wearable sensors *IEEE Int. Conf. on Data Mining Workshops* (IEEE)

Shmiel O, Shmiel T, Dagan Y and Teicher M 2009 Data mining techniques for detection of sleep arousals *J. Neurosci. Methods* **179** 331–7

Tsinalis O, Matthews P, Guo Y and Zafeiriou S 2016 Automatic sleep stage scoring with single-channel EEG using convolutional neural networks *CoRR* (arXiv: 1610.01683)

Varga B, Gorog M and Hajas P 2018 Using auxiliary loss to sleep arousal detection with neural network *Computing in Cardiology* vol **45**

Viitasalo J H T and Komi P 1977 Signal characteristics of emg during fatigue *Eur. J. Appl. Physiol.* **37** 111–21

Warrick P A and Homsi M N 2018a Ensembling convolutional and long short-term memory networks for electrocardiogram arrhythmia detection *Physiol. Meas.* **39** 1631

Warrick P and Homsi M 2018b Sleep arousal detection from polysomnography using the scattering transform and recurrent neural networks *Computing in Cardiology* vol **45**

Witten I H, Frank E, Hall M A and Pal C 2016 *Data Mining: Practical Machine Learning Tools and Techniques* (Cambridge, MA: Morgan Kaufmann)

Zeghidour N, Synnaeve G, Versteegh M and Dupoux E 2016 A deep scattering spectrum. Deep siamese network pipeline for unsupervised acoustic modeling *Proc. IEEE ICASSP* (IEEE) pp 4965–9

Zhang J and Wu Y 2018 Automatic sleep stage classification of single-channel EEG by using complex-valued convolutional neural network *Biomed. Eng./Biomed. Tech.* **63** 177–90