Arousal Detection in Obstructive Sleep Apnea using Physiology-Driven Features

Sandya Subramanian[1], Sourish Chakravarty[1], Shubham Chamadia[2]

[1]Massachusetts Institute of Technology, Cambridge, MA, USA (note - full address at end of paper)
[2]Massachusetts General Hospital, Boston, MA, USA

Abstract

Obstructive sleep apnea (OSA) is a condition in which a person repeatedly stops breathing during sleep due to closure of the upper airway, leading to a cycle of sleep fragmentation and intermittent hypoxia (oxygen deficiency). Conventional methods for detecting and quantifying OSA are largely based on physiological monitoring during sleep followed by manual labeling of sleep stages and arousals. Here there is scope for computerized methodologies that can efficiently and objectively perform this characterization of sleep.

As part of the CinC/Physionet 2018 challenge to automatically detect arousals in a large, expert-annotated sleep dataset, we extracted 27 spectral and time domain features, chosen for their physiological relevance, from the available training set and implemented two contrasting methods, Generalized Linear Model (GLM) and Random Forest (RF), to classify arousals and non-arousals.

We were able to achieve non-trivial classification accuracy, even in an imbalanced data set with far fewer arousals than non-arousals. This suggests that large machine learning problems can still benefit from physiology-informed feature selection, especially in the biomedical space.

1. Introduction

1.1. Obstructive Sleep Apnea

Obstructive sleep apnea (OSA) is a serious sleep disorder in which patients repeatedly stop breathing during deeper stages of sleep [1]. This is due to upper airway collapse, which leads to rapid desaturation of oxygen. The lack of oxygen (hypoxia) triggers the response of the sympathetic nervous system ("fight or flight" response), a natural response to stress in the body, which triggers recurring arousals to reopen the airway and restart breathing. Witnessed arousals, gasping, and choking arousals comprise only a small portion of all arousals, with most occurring without the conscious awakening of the patient at all. Other arousal types include spontaneous arousals, vocalizations, snores, bruxisms, periodic leg movements, Cheyne-Stokes breathing, or partial airway obstructions. Hypopneas are periods of abnormally slow or shallow breathing, and apneas are periods of no breathing.

While suspicion of OSA can arise from a variety of factors, including co-morbidities such as obesity [2], reported snoring or restless sleep at night or drowsiness during the day, or physical examination findings like narrowed airways, "kissing" tonsils, or abnormal Mallampati score, a diagnosis is usually confirmed in the setting of a sleep laboratory [3]. In such sleep laboratories, behavioral response and physiological signals are recorded from patients while they are asleep. Sleep experts then manually score different sleep stages as well as the presence of different types of arousals in epochs of 30 seconds for the whole night of sleep. In this context, automated detection of arousals could greatly reduce the time and cost required to diagnose OSA, in addition to reducing human error and developing an objective diagnosis scheme.

1.2. The Challenge and Related Dataset

The current work ensued from a competition conducted by Computing in Cardiology (CinC), whose dataset was made available on the Physionet website [4]. The competition aimed at detecting sources of arousal (non-apnea) during sleep using various physiological signals including electroencephalogram (EEG), electrooculogram (EOG), electromyogram (EMG), electrocardiogram (ECG), and blood-oxygen saturation (SpO2) that were each sampled at 200 Hz. This dataset consists of signals from 1983 subjects (994 for training and 989 for testing) that were recorded at Massachusetts General Hospital's (MGH) sleep laboratory dedicated for the diagnosis of sleep disorders. The dataset was also manually annotated for various stages of sleep (30 seconds intervals) as well as the presence of arousals events (hypopneas, snores, vocalization, etc.) by certified sleep technicians at MGH. The aim of the challenge was to correctly detect/classify the target arousal epochs, which included all arousal types except for apneic and hypopneic arousals. Specifically, the

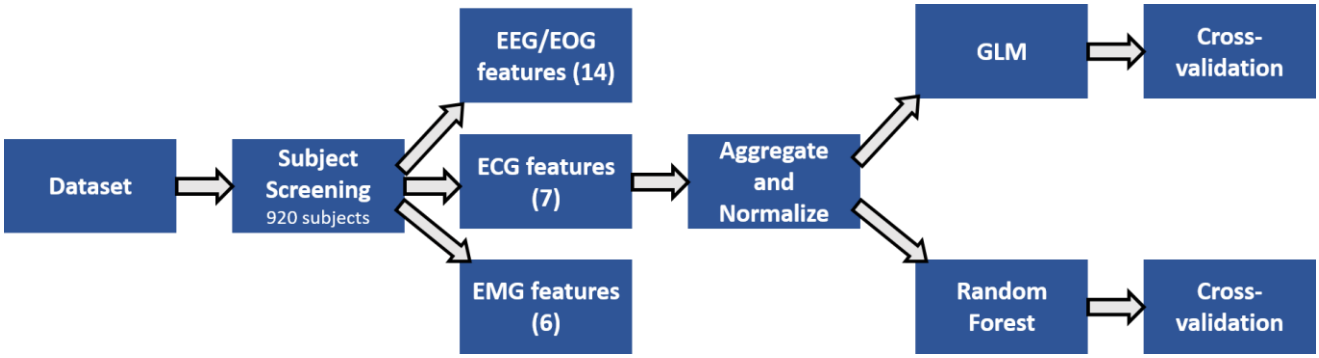

Figure 1 Overall schematic of the methods

scoring was based on how well the vectors of instantaneous probability of arousal predicted for each test subject was able to detect the non-apneic/non-hypopneic arousals. In this paper, however, we focus on the detection of arousals including apneic/ hypopneic and non-apneic/non-hypopneic arousals using two classification methods, the Generalized Linear Model (GLM) and the Random Forest (RF) (Fig. 1).

2. Methods & Results

We analyze all signals, sampled at frequency, $f = 200\,Hz$, with a time resolution Δ=5s since the target arousals themselves are often very short in duration (~2 sec). To down-sample the arousal states (0: non-arousal, +1: non-apneic/non-hypopneic arousal, -1: apneic/hypopneic arousals), we use a "majority vote" polling method to set the arousal state, y_k, based on the mode of the trinary observations within the kth window of size Δ . When missing data scenario is encountered at any instant in any of the signals, the corresponding Δ window is omitted. In each modality, features are chosen based on physiological relevance to detection of arousals in the context of sleep apnea. Since the focus for this paper is detection of all arousals, the -1's are converted to +1's in the arousal state vector prior to model fitting.

2.1. EEG and EOG Features

For every window of size Δ, we extract features from the Multitapered Power Spectral Density (PSD) [5], F_j, for the frequency bin, ω_j (in Hz) with constant width W and using a stationary time window = Δ, time-half-bandwidth product = 2, and number of tapers = 3. We perform the following regression, $log_{10}(F) = b - c\,log_{10}(\omega)$, for the kth window of size Δ, to identify parameters, b_k and c_k, that characterize the background '1/f' decay [6]. Then, using the residuals, $log_{10}\left(R_k(\omega_j)\right) = log_{10}(F_j) - (b_k - c_k log_{10}\omega_j)$ we calculate the following parameters, $\delta_k = W\sum_{0<\omega_j\leq 4} R_k(\omega_j)$, $\theta_k = W\sum_{4<\omega_j\leq 8} R_k(\omega_j)$, $\alpha_k = W\sum_{8<\omega_j\leq 14} R_k(\omega_j)$, $\beta_k = W\sum_{14<\omega_j\leq 30} R_k(\omega_j)$ and $\gamma_k = W\sum_{30<\omega_j\leq 55} R_k(\omega_j)$. Finally, for each subject, the features are rescaled as, $z_k = (x_k - \mu)/\sigma$, where x_k refers to any of the 7 fields, $[b_k, c_k, \delta_k, \theta_k, \alpha_k, \beta_k, \gamma_k]$, extracted for any of the 6 EEG and 1 EOG channels and μ and σ, respectively, correspond to the subject-specific mean and standard deviation of the field associated with x_k. Our choice of the frequency bands for the EEG is inspired by those used for tracking sleep-stages [7]. Here our implicit assumption is that the sleep stages can be correlated to arousal states. For computational efficiency, we decided to use 1 EEG channel based on our initial exploratory data analysis of the features across 6 EEG channels. The empirical distributions of the features corresponding to arousal and non-arousals from 1 EEG channel and 1 EOG channel are illustrated in Fig. 2.

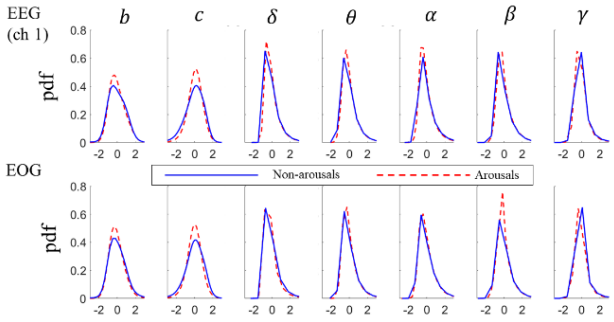

Figure 2 Kernel density plots of EEG and EOG features for all arousals and non-arousals in the training set.

2.2. ECG Features

The ECG is pre-processed by extracting R peaks using the Pan-Tompkins algorithm [8]. RR intervals that are too long (> 2 seconds) or too short (< 0.3 seconds) are corrected with the addition or removal of R peaks respectively. After the extraction of R peaks, four time-domain ECG features are computed for each window of size Δ as follows: (1) mean of all RR intervals within that

window, (2) standard deviation of all RR intervals within that window, (3) root mean square of the difference between consecutive RR intervals within that window, and (4) proportion of differences between consecutive RR intervals greater than 0.05 seconds within that window.

Then the multitapered power spectral density is computed for each stationary time window = Δ, time-half-bandwidth product = 3, and number of tapers = 5. Three more frequency-domain features are computed as follows: (5) total power between 0.04 and 0.15 Hz (low frequency), (6) total power between 0.15 and 0.40 Hz (high frequency), and (7) the ratio of low frequency to high frequency power within that window. All ECG features are heartrate variability measures, which are good indicators of autonomic tone. Since the autonomic nervous system is strongly affected during arousals ("fight-or-flight" response), we hypothesize that these features would be useful. The empirical distributions of the features extracted from the ECG are illustrated in Fig. 3.

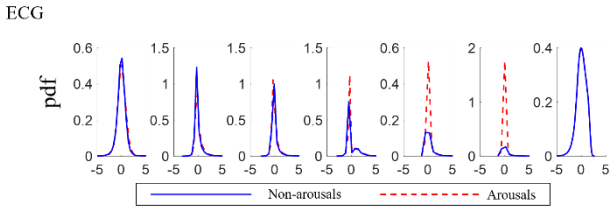

Figure 3 Kernel density plots of ECG features for all arousals and non-arousals in the training set

2.3. EMG Features

The dataset contained three different EMG channels: chin, chest, and abdomen. Chin EMG contained mainly high frequency activity relating to jaw clench, while chest and abdomen contained much lower frequency sinusoidal activity correlating with the rise and fall of the chest during breathing. Two features are extracted for each channel as follows: (1) total power in each window of size Δ from 0-100 Hz for chin and 0-5 Hz for chest and abdomen, and (2) difference in total power between consecutive windows for each channel. All spectral density estimates are computed using the multitapered power spectral density, time-half-bandwidth-product = 3, number of tapers = 5. The empirical distributions of the EMG features are illustrated in Fig. 4.

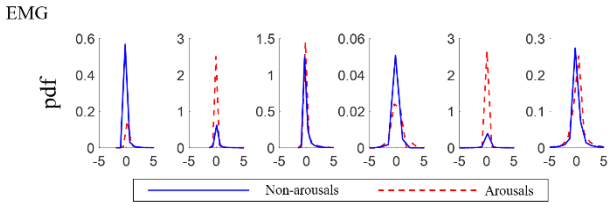

Figure 4 Kernel density plots of EMG features for all arousals and non-arousals in the training set

2.4. Classification schema

We set up the analysis scheme for both classification methods, GLM and RF, as follows. As mentioned earlier, we group non-target arousals (apneic and hypopneic arousals) with target arousals. Arousal states are down-sampled as previously described to 5-second windows. With such coarse-grained arousal observations and 27 features extracted at the same time resolution, the models can be learnt from a training data set. These predictive models can be implemented on any test case to estimate a vector of instantaneous probabilities of arousal. Once such predictions are computed for every 5 second window, they can be expanded to yield prediction probability vectors at the sampling rate (200 Hz) by simply assigning the prediction computed for each window to all 1000 "samples" within the 5 second window. When tested on datasets where arousal observations are available, metrics such as the area under the receiver operating curve (AUROC) and the area under the precision recall curve (AUPRC) can be computed. Such metrics are useful to analyze the classification abilities of the chosen methods.

Within the training dataset provided by the competition, we identified 74 (out of 994) subjects as outliers in the feature space and discarded them from the model fitting and validation steps. Across the remaining 920 subjects, the mean proportion of arousals compared to non-arousals is found to be 0.201.

2.5. A naïve Generalized Linear Model

The first method we used for classification was a GLM of the form

$$E\left[log\left(\frac{p}{1-p}\right)\right] = a_0 + \sum_{p=1}^{P} a_p x_p$$

to classify arousals vs. non-arousals. Here, P denotes the number of features used, p is the instantaneous probability of an arousal, $E[\,]$ denotes the expectation operator and x_p indicates the pth feature. To analyze the out-of-sample error from GLM, we use a cross-validation scheme on 920 subjects from the training set by partitioning it into 10 folds with 92 subjects in each. Each fold consisted of around 5000-6000 data points. For every fold, we fit the GLM on the training data from the other 9 folds, and then apply the estimated model to predict the arousal probabilities for each subject from the held-out fold. Using these predicted arousal probabilities and available ground truth arousal data, we calculate the AUROC and AUPRC. The GLM training and testing are conducted using pre-defined MATLAB functions *glmfit()* and *glmval()*.

2.6. Random Forest

The second method we used for cross-validation is

Random Forest. Like the GLM, the training data set is partitioned into the same 10 folds with 92 subjects in each. A random forest of 100 weak learners is trained on each fold, where each decision tree is constrained to having no more than seven splits. The LogitBoost algorithm is used to boost accuracy for all forests. The pre-defined MATLAB function *cfitensemble()* was used to train all forests. Then for each of the 920 subjects, predictions for each window are computed as the average of the predictions yielded by the nine forests not trained on that subject's data (the other nine folds). The AUROC and AUPRC from both GLM and RF are compared against baseline in Table 1.

Table 1 Summary of final results using GLM and RF

Method	**AUROC**	**AUPRC**
Baseline	0.5	0.201
GLM	0.601	0.263
Random Forest	0.694	0.364

3. Discussion & Conclusion

There are several points worthy of note from this work. First, we used principled tools to extract features from 920 subjects and analyze them. This is a difficult classification problem, as evidenced by the fact that most of the features, when viewed individually, do not demonstrate clear separation between arousal and non-arousals, and any visible separation is often in spread of distribution rather than center. However, we are still able to achieve non-trivial classification power with just 27 features, likely because the joint distributions of these physiologically-relevant features from arousals and non-arousals are statistically distinguishable. This study, therefore, illustrates the potential of physiology driven feature selection for machine learning problems in biomedicine and disease.

Second, the physiological relevance of the features can also be used for other classification questions, such as sleep staging. For example, sleep stages are often distinguished based on their EEG spectral signatures, which we also used as features in this work. Further extensions of this work could be in sleep stage determination for patients with obstructive sleep apnea.

Finally, we contrast two different classification methodologies that come from completely different schools of thought. GLM assumes a parametric distribution for the data, while RF is a fully non-parametric method that makes no such assumptions, but relies on a deterministic framework for classification. While working on the same set of features, RF appears to work better then GLM (Table 1), indicating that the chosen model for GLM might be too restrictive. However, that RF is able to yield AUROC of about 0.7 and AUPRC of 0.364 indicates that the chosen features are relevant to discern arousal events from non-arousal. Note that neither of these methods is black box; in both cases, potentially valuable information regarding which features are important can be extracted post-hoc to help advance our understanding of the disease itself. Future work could involve comparison to black box methods, such as artificial neural networks.

Acknowledgements

We would like to thank Leon Chlon and Pegah Kahaliardabili for their ideas and assistance. We would also like to acknowledge the resources given to us by the Brown Lab, Akeju Lab, MIT, and MGH.

References

[1] Downey III R, et al. "Obstructive Sleep Apnea: Practice Essentials, Background, Pathophysiology." *Medscape*, 9 Jan. 2018, emedicine.medscape.com/article/295807-overview.

[2] Ho ML, Brass SD. Obstructive sleep apnea. Neurol Int. 2011;3:e15

[3] McNicholas WT. Diagnosis of obstructive sleep apnea in adults. Proc Am Thorac Soc 2008;5:154–160

[4] PhysioNet. PhysioNet/Computing in Cardiology Challenge 2018. http://www.physionet.org/challenge/2018/

[5] Thomson DJ. Spectrum estimation and harmonic analysis. Proceedings of the IEEE, 1982 Sep;70(9):1055-96.

[6] Haller M, Donoghue T, Voytek B et al. Parameterizing Neural Power Spectra, BioRxiv, 2018, 299859.

[7] Prerau MJ, et al. "Tracking the sleep onset process: an empirical model of behavioral and physiological dynamics." *PLoS computational biology* 10.10 (2014): e1003866.

[8] Pan J, Tompkins WJ. A Real-Time QRS Detection Algorithm. IEEE Transactions on Biomedical Engineering, 1985 Mar; 32(3): 230-36.

Address for correspondence.

Sandya Subramanian
77 Massachusetts Ave. 46-6057A
Cambridge, MA 02139
sandya@mit.edu