You Snooze, You Win: the PhysioNet/Computing in Cardiology Challenge 2018

Mohammad M Ghassemi[1], Benjamin E Moody[1], Li-wei H Lehman [1], Christopher Song [2], Qiao Li[3], Haoqi Sun[5], Roger G Mark[1], M Brandon Westover[5], Gari D Clifford[3,4]

[1] Institute for Medical Engineering & Science, Massachusetts Institute of Technology, USA
[2] Malone Center for Engineering in Healthcare, Johns Hopkins University, USA
[3] Department of Biomedical Informatics, Emory University, USA
[4] Department of Biomedical Engineering, Georgia Institute of Technology, USA
[5] Department of Neurology, Massachusetts General Hospital, USA

Abstract

The PhysioNet/Computing in Cardiology Challenge 2018 focused on the use of various physiological signals (EEG, EOG, EMG, ECG, SaO_2) collected during polysomnographic sleep studies to detect sources of arousal (non-apnea) during sleep. A total of 1,983 polysomnographic recordings were made available to the entrants. The arousal labels for 994 of the recordings were made available in a public training set while 989 labels were retained in a hidden test set. Challengers were asked to develop an algorithm that could label the presence of arousals within the hidden test set. The performance metric used to assess entrants was the area under the precision-recall curve. A total of twenty-two independent teams entered the Challenge, deploying a variety of methods from generalized linear models to deep neural networks.

1. Introduction

The PhysioNet/Computing in Cardiology Challenge is a competition centered around the creation of open-source software solutions for complex physiological signal processing problems. In 2018, we challenged entrants to develop automated techniques for the detection of non-apnea sleep arousals. To facilitate the development of their algorithms, we provided a variety of physiological signals, collected during polysomnographic (PSG) sleep studies.

Sleep is critical to health and well-being. Inadequate or poor quality sleep is associated with a wide range of negative outcomes, including: impaired cognitive and motor function, irritability [1], obesity [2], and depression [3, 4].

"Arousals" are brief intrusions of wakefulness into sleep, after which sleep resumes [5]. Spontaneous arousals are a normal feature of the sleeping brain [6]. However, arousals also occur in response to sleep-disturbances, and when excessive can cause harm. Arousals are often the result of obstructive sleep apnea or hypopnea events. Arousals may also be respiratory effort-related (RERA) or due to more minor/subtle temporary obstructions that are not severe enough to meet the criteria for apneas/hypopneas. Other causes of pathological arousals include teeth grinding (bruxism), muscle jerks (including "periodic limb movements of sleep"), pain, insomnia, and even snoring. Sleep fragmentation – frequent interruption of sleep by arousals – results in daytime sleepiness, degraded cognitive performance, and generally decreases the ability of sleep to perform its recuperative functions [7–9].

It follows that improving the quality of sleep could be used to improve a range of societal health outcomes, more generally. Of course, the treatment of sleep disorders is necessarily preceded by the diagnosis of sleep disorders. Traditionally, such diagnoses are developed in sleep laboratory settings, where PSG, audio, and videography of sleeping subject may be carefully inspected by sleep experts to identify potential sleep disorders.

2. Challenge Data

2.1. Data Source

A total of 1,983 PSG recordings were provided by the Massachusetts General Hospital's (MGH) Sleep Lab in the Sleep Division together with the Computational Clinical Neurophysiology Laboratory, and the Clinical Data Animation Center. The Partners Institutional Review Board approved retrospective analysis of the MGH dataset without requiring additional consent.

The technicians captured the PSG following the AASM standards. There were thirteen signals including six channels of electroencephalography (EEG) at F3-M2, F4-M1, C3-M2, C4-M1, O1-M2, and O2-M1 based on the International 10/20 System; electroculography (EOG) on the

Clinical Feature	Total	Training	Test
Sample size	1,983	994	989
Age	55 (14.4)	55 (14.3)	55 (14.4)
Gender (% male)	65	67	63
BMI	33 (7.6)	33 (7.8)	33 (7.5)
AHI	19 (14.4)	19 (14.6)	18.9 (14.4)
ESS	8.6 (5.3)	8.5 (5.3)	8.7 (5.3)
Recording time (h)	7.7 (0.7)	7.7 (0.7)	7.7 (0.7)
Time in bed (h)	7.5 (0.7)	7.5 (0.7)	7.5 (0.7)
Sleep time (h)	6.2 (1.2)	6.2 (1.1)	6.1 (1.2)
Drug Use (%)			
Hypertension	40.9	41.0	40.6
Sleep aids	28.3	29.0	27.8
Antidepressant	26.1	25.7	26.5
Neuroactive	19.1	20.8	17.5
Benzodiazepine	16.1	16.9	15.4
Diabetic	11.7	11.9	11.5
Opiate	7.4	8.1	6.7
Antihistamine	4.8	4.8	4.8
Stimulant	4.7	3.9	5.5
Neuroleptic	4.2	4.5	3.8
Herbal	4.2	4.3	4.0
Reason for visit (%)			
Diagnostic	41.8	41.16	42.47
Split night CPAP	38.35	37.95	39.03
All night CPAP	19.85	20.88	18.5

BMI: Body Mass Index; AHI: Apnea-Hypopnea Index;
ESS: Epworth Sleepiness Scale;
CPAP: Continuous Positive Airway Pressure

Table 1. Clinical characteristics of the MGH dataset.

left side (EEG and EOG referenced to the contralateral ear lobe); electromyography (EMG) was measured at the chin; two channels of respiration signal from the abdomen and chest; airflow and oxygen saturation (SaO_2); and one ECG channel recorded below the right clavicle near the sternum and over the left lateral chest wall. All signals except the SaO_2 were measured with a sampling frequency of 200Hz. The SaO_2 was upsampled using *sample and hold* to 200Hz to synchronize samples. All signals were measured in microvolts.

A total of 1,983 PSG recordings were made available to the entrants. The arousal labels for 994 of the recordings were made available in a public training set while 989 labels were retained in a hidden test set. Since apnea is one of the causes of arousal, the dataset was partitioned to ensure a uniform distribution of AHIs in both sets (Kolmogorov-Smirnov test p-value 0.97). There were no subjects in common between the training and test sets. The test set labels were maintained to be private during the Challenge, and will remain so to enable follow-up works. Characteristics of the dataset are summarized in Table 1.

Clinical Feature	Total	Training	Test
Time spent in sleep stage (%)			
Wake	29.3	28.0	31.0
NREM 1	19.5	19.6	19.0
NREM 2	51.3	51.0	51.7
NREM 3	13.8	14.0	13.8
REM	15.3	15.5	15.2
Number of target arousals			
Bruxism	–	30	–
Cheyne-Stokes breathing	–	3	–
Hypoventilation	–	4	–
Noise	–	1	–
Partial airway obstruction	–	11	–
PLM	–	36	–
RERA	–	43,822	–
Snoring	–	28	–
Spontaneous	–	70	–
Number of non-target arousals			
Hypopnea	–	56,936	–
Central apnea	–	22,763	–
Mixed apnea	–	2,641	–
Obstructive apnea	–	32,547	–

PLM: Periodic leg movement
RERA: Respiratory effort-related arousals

Table 2. Sleep/arousal characteristics of the MGH dataset.

2.2. Expert Labeling

In total, seven scorers annotated the dataset, but with one scorer per PSG. The EEG signals were scored in nonoverlapping 30-second epochs according to the AASM standards as one of five stages: wake (W), rapid eye movement (REM), non-REM stage 1 (N1), non-REM stage 2 (N2), and non-REM stage 3 (N3). Subject waveforms were also annotated for the presence of arousals that interrupted their sleep. 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.

3. Challenge Objective

The goal of the Challenge was to use information from the available signals to correctly classify target arousal regions. For the purpose of the Challenge, target arousals were defined as regions where either of the following conditions were met:

- From 2 seconds before a RERA arousal begins, up to 10 seconds after it ends or,

• From 2 seconds before a non-RERA, non-apnea arousal begins, up to 2 seconds after it ends.

Regions falling within 10 seconds before or after a subject awoke, had an apnea arousal, or a hypopnea arousal were not scored.

4. Scoring

Each team was asked to submit a complete, working implementation of their algorithm that could be run in the Challenge sandbox environment[1]. For each test subject, entrants were also required to submit a vector providing the probability of the target arousal, at the sample level.

During the official phase of the Challenge (April through August, 2018), each team could submit up to two entries for scoring. When an entry was submitted, the automated scoring system calculated a "provisional" score, based on the entry's performance on a *subset* of the test data, to provide a rough form of feedback to the authors. At the conclusion of the Challenge, final scores were calculated based on the complete test set. If a team submitted two entries, they were asked to choose one of the two to be considered as their final entry.

Each competing team's final algorithm was graded for its binary classification performance on target arousal and non-arousal regions, as measured by the area under the precision-recall curve (AUPRC). Precision (p_j) and recall (r_j) were defined as follows:

$$p_j = \frac{|A \cap P_j \cap \overline{N}|}{|P_j \cap \overline{N}|}$$

$$r_j = \frac{|A \cap P_j \cap \overline{N}|}{|A \cap \overline{N}|}$$

where N indicates the set of *non-scored* samples, A indicate the set of *target arousal* samples, and P_j indicates the set of samples for which the predicted arousal probability was at least $\frac{j}{1000}$. The area under the curve (and the team's final score) was calculated accordingly:

$$AUPRC = \sum_{j,\, |P_j \cap \overline{N}| \neq 0} p_j (r_j - r_{j+1})$$

Note that this is the gross AUPRC (i.e., for each possible value of j, the precision and recall are calculated for the entire test database), as opposed to averaging the AUPRC for each record. More information on the Challenge scoring mechanism and rules can be found at [10].

If the probability vector produced by the entry was not of the correct length, it was truncated or padded with zeroes accordingly. If the entry failed to run for a particular record, it was treated as if it had produced a vector of all zeroes.

[1] https://physionet.org/challenge/sandbox/

Rank	Entrant	AUPRC
1	Howe-Patterson, Pourbabaee & Benard	0.54
2	Kristjánsson, Þráinsson, Ragnarsdóttir, Marinósson, Gunnlaugsson, Finnsson, Jónsson, Helgadóttir, & Ágústsson	0.45
3	He, Wang, Liu, Zhao, Yuan, Li, & Zhang	0.43
4	Varga, Görög, & Hajas	0.42
5	Patane, Ghiasi, Scilingo, & Kwiatkowska	0.40
6	Miller, Ward, & Bambos	0.36
6	Warrick & Homsi	0.36
8	Bhattacharjee, Das, Choudhury, & Banerjee	0.29
8	Szalma, Bánhalmi, & Bilicki	0.29
10	Parvaneh, Rubin, Samadani, Prakash, & Katuwal	0.21
11	Plešinger, Nejedly, Viscor, Andrla, Halámek, & Jurák	0.20
12	Zabihi, Rad, Särkkä, Kiranyaz, Katsaggelos, & Gabbouj	0.19
13	Schellenberger, Shi, Mai, Wiedemann, Steigleder, Eskofier, Weigel, & Kölpin	0.14
14	Li, Cao, Zhong, & Pan	0.10
15	Jia, Yu, Yan, Zhao, Xu, Hu, Wang, & You	0.10
16	Shen	0.07
	Unofficial entries	
–	*Li & Guan* †	0.55
–	*Bilal, Khan, Khan, Qureshi, Saleem, & Kamboh* †	0.15
–	*Wang, Wang, & Li* †	0.07

Table 3. Final scores for the 19 teams in the Challenge. † denotes unofficial entries.

5. Results

During the official period of the competition, a total of 624 entries were submitted by 34 distinct PhysioNet users. Of the 624 entries, the majority (82%) were submitted as dry-runs to test code compatibility with the sandbox environment. A total of 37 valid entries were submitted for scoring by 24 distinct users; 19 of these entries (one per team) qualified for final scoring. The final entries carried the following open source software licenses: MIT X11 License ($n = 8$), GNU General Public License version 3 ($n = 10$), and GNU GPL version 2 ($n = 1$). Table 3 lists the final scores on the test set. We rounded to two decimal places for awarding prizes.

6. Conclusions

The excellent performance of the first-place entry ($AUPRC = 0.54$), as well as the unofficial top score, indicate that automated arousal detection is realizable. However, the large variance in performances across entrants (mean 0.28, and ranging from 0.07 to 0.55) indicates that arousal detection is a challenging problem. Deep neural-network approaches have gained significant interest and traction in recent years, and this year's Challenge was no exception. Thirteen of the 22 final entries used neural networks as a component of their arousal detection algorithms. Among the top ten performers, eight used neu-

ral networking approaches, but neural networks were also used regularly among the bottom ten entries (n=4). This variance in performance validates observations from the greater community of scholars about neural network approaches: they are powerful when used correctly. The open source nature of the Challenge should be helpful in this regard, as all models may be downloaded, inspected, and re-purposed for related, or unrelated problems in the general domain of physiological signal processing (e.g., via transfer learning).

The 2018 Challenge was subject to several limitations. First, although a time limit of 1.2×10^{13} CPU instructions (about 3.5 hours) was enforced, our scoring approach did not explicitly penalize or reward algorithms for their computational (in)efficiency. This unfairly equates an entry that requires two hours of computation, with one that takes only two minutes for the same AUPRC performance. In reality, computationally inefficient approaches are inferior to those with similar classification performances that require less time. Future challenges may overcome this important limitation by explicitly including computational efficiency in the scoring function.

Another limitation of the 2018 Challenge was the use of a performance metric that disregards model calibration. Models that are well-calibrated may be used descriptively, while classification-only metrics (i.e., AUPRC) are limited to prescriptive use. That is, a well-calibrated model may state the probability of an important arousal event with a confidence interval. That may be more useful, in practice, than a binary indicator of the arousal. This important aspect of model performance is not accounted for by the AUPRC. Future challenges may benefit from considering a measure of statistical calibration.

Several of the entries exhibited non-deterministic behavior. That is, they produced similar, albeit not perfectly identical, classifications of arousal segments when applied multiple times to the same records. This limitation may have impaired our ability to perfectly characterize the algorithm's performance, and prohibits us from making *strong* guarantees about future performance on new data.

Having addressed the limitation of the 2018 Challenge, we believe that the data and algorithms produced and made publicly available as part of the Challenge represent an important contribution to the physiological signal processing community specifically, and greater scientific community more generally. There are three key features of the PhysioNet Challenge that distinguish it from other data science competitions (e.g., Kaggle). First is our focus on collecting, and publicly releasing, well-curated novel datasets in the domain of physiology. Second, and more importantly, is the open-source spirit (and formal requirement) of the Challenge. Entrants are not only required to submit code that runs in an external sandbox ecosystem (ensuring reproducibility), but must also document their approaches as part of an academic paper submitted to the annual conference: Computing in Cardiology. Third, competitors must attend a public forum and verbally defend their work. Together, these requirements ensure that open-source, well-documented, reproducible software is developed and distributed every year as a direct consequence of the PhysioNet Challenge.

References

[1] Pilcher JJ, Huffcutt AI. Effects of sleep deprivation on performance: A meta-analysis. Sleep 1996;19(4):318–326.

[2] Ogilvie RP, Patel SR. The epidemiology of sleep and obesity. Sleep Health 2017;3(5):383–388. ISSN 2352-7218.

[3] Nutt D, Wilson S, Paterson L. Sleep disorders as core symptoms of depression. Dialogues in clinical neuroscience 2 2008;10:329–36.

[4] Lee M, Choh A, Demerath E, Knutson K, Duren D, Sherwood R, Sun S, Chumlea W, Towne B, Siervogel R, Czerwinski S. Sleep disturbance in relation to health-related quality of life in adults: The fels longitudinal study. Journal of Nutrition Health and Aging 2009;13(6):576–583.

[5] ASDA E. arousals: scoring rules and examples: a preliminary report from the sleep disorders atlas task force of the american sleep disorders association. Sleep 1992; 15(2):173–184.

[6] Boselli M, Parrino L, Smerieri A, Terzano MG. Effect of age on eeg arousals in normal sleep. Sleep 1998;21(4):361–367.

[7] Bonnet MH. Effect of sleep disruption on sleep, performance, and mood. Sleep 1985;8(1):11–19.

[8] Bonnet MH. Performance and sleepiness as a function of frequency and placement of sleep disruption. Psychophysiology 1986;23(3):263–271.

[9] Stepanski E, Lamphere J, Roehrs T, Zorick F, Roth T. Experimental sleep fragmentation in normal subjects. International journal of neuroscience 1987;33(3-4):207–214.

[10] Ghassemi M, Moody B, Song C, Haoqi S, Westover B, Clifford G. You Snooze, You Win: the PhysioNet/Computing in Cardiology Challenge 2018, 2018 (accessed Sept 3, 2018). `http://physionet.org/challenge/2018`.

Acknowledgments

Funding was from the National Institutes of Health, grant R01-GM104987. We are also grateful to Mathworks and Computing in Cardiology for sponsoring the competition prize money and software licenses.

Address for correspondence:

Gari Clifford; gari@physionet.org
WMRB, 1639 Pierce Dr NE, Atlanta, GA 30322