Evaluating Convolutional and Recurrent Neural Network Architectures for Respiratory-Effort Related Arousal Detection During Sleep

Ali Shoeb, Niranjan Sridhar

Verily Life Sciences, South San Francisco, California

Abstract

This work evaluates the performance of convolutional and recurrent neural networks on the task of detecting Respiratory Effort-Related Arousals (RERAs). Feature time-series were extracted from EEG, EOG, CHIN, CHEST, ABDOMINAL, AIRFLOW, SaO2, and ECG and normalized on a per-subject basis. Next, multi-timescale windows from these time-series were associated with the presence or absence of RERA during the window forming the data for model training. More than 1 million RERA-windows and 17 million no-arousal windows were used for model training, and more than 200K RERA-windows and 4 million no-arousal windows were used for testing and validation. Google Cloud ML Engine was used to select model hyperparameters using the validation data. The model with the best hyperparameter combination evaluated on the test set achieved an AUC-ROC score of 0.916 and AUC-PR score 0.573.

1. Introduction

Sleep disruption has been correlated with a wide array of negative repercussions on health [1]. Sleep-related breathing disorders are chief among the causes of inadequate and fragmented sleep. These disorders involve complete (apnea), partial (hypopnea), or subtle (RERA) obstruction of the upper airways which results in episodic asphyxia and interruption of normal sleep [1].

The ability to manage sleep-related breathing disorders depends on the ability to detect and distinguish between them [2]. The American Academy of Sleep Medicine (AASM) provides clear guidelines for establishing whether respiratory disturbances are apneas or hypopneas [3]. However, both the subtlety and variability of RERAs renders defining exact criteria a challenge.

Currently, the AASM defines a RERA to be a "sequence of breaths lasting at least 10 seconds characterized by either increasing respiratory effort or flattening of the inspiratory portion of the nasal pressure or PAP device flow waveform leading to arousal from sleep when the sequence of breaths does not meet criteria for an apnea or hypopnea event" [3].

Given this definition, the agreement among raters for identifying RERA events is substantial but imperfect. Using non-invasively measured thoracoabdominal belt waveforms intra-rater and inter-rater agreement (kappa score) is 0.80 and 0.85 respectively. These are lower than the intra-rater and inter-rater agreement of 0.91 and 0.89 using esophageal manometry; the invasive and poorly tolerated gold-standard method for detecting RERAs [4].

The combination of imperfect RERA event scoring, impractical gold-standard measurements, and the time-consuming process of expert sleep scoring [5], establishes the need for automated methods that aid in the annotation of sleep studies. This work focuses on the development and evaluation of RERA event detectors that use neural networks to process noninvasive signals commonly collected in polysomnography studies.

2. Dataset

2.1 Data Description

The dataset used in this work was made available through the PhysioNet 2018 Challenge [7]. The dataset consists of polysomnography recordings from 994 subjects. The recordings include EEG, EOG, EMG, EKG, and SaO2 signals sampled at 200 Hz and annotations supplied by certified sleep technologists.

The majority of RERA events in the dataset (99.7%) have durations shorter than 2 minutes, and the mean event duration is 30 +/- 15 seconds. The median of mean inter-arrival time of RERA events aggregated per subject is 15 min. Furthermore, RERA events are asymmetrically distributed across sleep stages in the dataset. The sleep stages N1, N2, N3, and REM account for 24%, 59%, 4.4%, and 12% of RERA events respectively.

2.2 Train, Test, Validation Split

The 994 polysomnography recordings were randomly split into a training set of 793 recordings, 97 test recordings, and 102 validation recordings. The training recordings generated approximately 17 million no-arousal and 1 million RERA tensors; the validation and test

 ISSN: 2325-887X DOI: 10.22489/CinC.2018.284

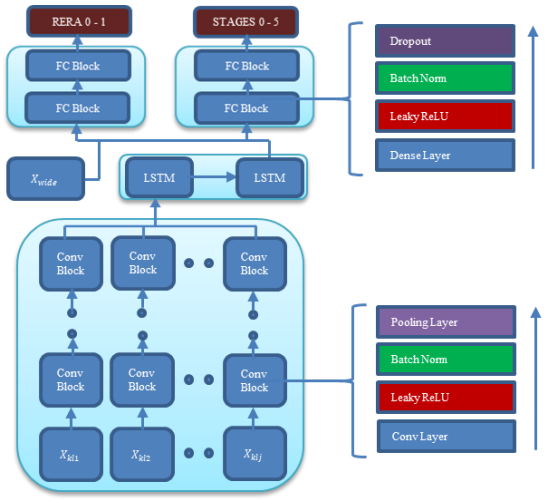
recordings had approximately 2.1 million no-arousal and 142K RERA tensors.

3. Methods

3.1 Neural Network Architecture

The network architectures evaluated are variations of the architecture illustrated in Figure 1. All networks are binary classifiers determining whether an observation is consistent with the absence of an arousal or the presence of a RERA. Therefore, the network's response to other arousal types (e.g. apneas) cannot be easily predicted. Such arousals have stereotyped manifestations enabling their detection with other methodology [6].

The networks receive a collection of tensors X_{kl} $k = 1 \dots 4,\ l = 1 \dots 3$ constructed from features derived from polysomnography signals (Section 3.2). Each tensor X_{kl} has dimensions $B \times W_k \times C_l$ where B is the batch size. The width of the tensor $W_k = T_k \times fs$, where fs is the sampling rate and T_k is the temporal window which can be $T_1 = 30$, $T_2 = 90$, $T_3 = 120$ or $T_4 = 180$ seconds. The number of channels C_l corresponds to our choice of feature groups: 1) all features together $C_1 = 20$, 2) all features separate $C_2 = 1$ or 3) splitting the features into 4 groups of features $C_3 = 5$.



Each tensor X_{kl} in this collection is passed into a separate convolutional tower. A tower is composed of a variable number of convolutional blocks each containing: 1) variable number of convolutional layers 2) optional batch-normalization layer 3) max or average pooling 4) L1/L2 regularization 5) optional drop-out layer.

Let the shape of the tensors produced by tower j be $B \times W'_j \times C'_j$ where B denotes batch size, W'_j denotes the width and C'_j denotes the number of channels of the output tensor. When using recurrent layers, the tower outputs can be fused to form a single tensor to be fed into a single RNN or fed into separate RNNs. The RNN stack can have 0 to 3 optionally bidirectional LSTM layers. The RNN produces an output shaped $B \times W'_j \times R_j$, where $R_j = R$, the size of the LSTM cell output ($R_j = W'_j C'_j$ if no RNN layers are used). This tensor is then reduced to $B \times R_j$ by taking the mean of the time axis. The outputs of the recurrent layers are then concatenated to form the tensor X_{deep} of shape $B \times (\sum_j R_j)$.

We also added some optional "wide" features. Specifically we tried the power spectral densities of all night time-series and a one-hot encoding of the most prevalent sleep stage in the label window. These features are concatenated and passed through a variable number of fully-connected (FC) blocks each containing: 1) fully-connected layer 2) optional batch-normalization 3) drop-out layer. Let the output of the final block be X_{wide}.

Finally, the tensors X_{deep} are optionally concatenated with X_{wide} and then processed by a variable number of FC blocks before the final softmax layer. The label for training was the most prevalent arousal state in the central 2, 10 or 30 seconds of the feature window. In addition, we added an optional task to predict the sleep stage label.

3.2 Feature Time-Series

Different feature time-series are extracted depending on the identity of the polysomnography signal.

EEG, EOG, and Chin: The feature time-series extracted from these channels consisted of spectral energies in the delta (0.5-3 Hz), theta (3-8 Hz), alpha (8-13 Hz), beta (13-25 Hz), and gamma bands (25-50 Hz) computed over a 10-second window.

CHEST, ABOMINAL, AIRFLOW: The time-series extracted from these respiratory channels include breath rate, breath width, breath amplitude, inspiratory slope, expiratory slope, and inter-breath intervals.

SaO_2: The time-series derived from the pulse-oximetry channel was the rolling mean over a 10-second window.

ECG: Heart rate, inter-beat intervals and R-wave amplitude time-series from the electrocardiogram.

In addition, a rolling variance over a 10-second window was also computed for all raw and ECG derived signals. All the above signals were sampled at multiple sampling rates between 1-5 Hz.

3.3 Hyperparameter Search

Tuning experiments involved training more than 300 model instances with randomly chosen hyperparameters and ranking them according to test set performance. Hyperparameters include: 1) subset of polysomnography signals used 2) combination of tensors used 3) usage of wide features 4) number of convolutional layers, blocks, filters and kernel sizes 5) joint or separate convolutional towers 6) usage of batch-normalization 7) max or average

pooling 8) size and number of recurrent layers 9) size and number of the FC layers 10) dropout, regularization, learning rates and weighting of positive examples.

Certain hyperparameters such as the width of the label window, usage of the optional stage prediction task and feature time-series sampling rates were tested and compared using similar but separate experiments.

4. Results

4.1 Overall Results

The best model achieved an AUC-ROC (Area Under Receiver Operating Characteristic Curve) score of 0.916, and AUC-PR (Area Under Precision Recall Curve) score of 0.573 considering all validation tensors. For reference, a model that predicts the majority class has AUC-PR of 0.

4.2 Hyperparameter Search

The best model architecture used the input tensors X_{33}, i.e 120 seconds feature windows in 4 groups, at a sampling rate of 5 Hz and a label definition of the central 2 sec of the window. The feature tensors were processed using 4 separate towers, each with 2 convolutional blocks, 4 layers per block, kernel size of 3 and 64 filters for each layer. Batch norm was not used and max pooling was used. Two recurrent layers without dropout were used. No wide features or fully-connected layers were applied before the classification layer. Finally, L1 regularization scale was 0.5, RERA examples were weighted twice as much as normal examples and both RERA and stages prediction losses were trained together.

4.3 Subject-wise Results

The best model's AUC-PR score varied widely across validation recordings (min 0.007, max 0.92) and was positively correlated (pearson correlation: 0.39 p-value 9.6e-5) with a subject's respiratory disturbance index (RDI) as shown in Fig 2.

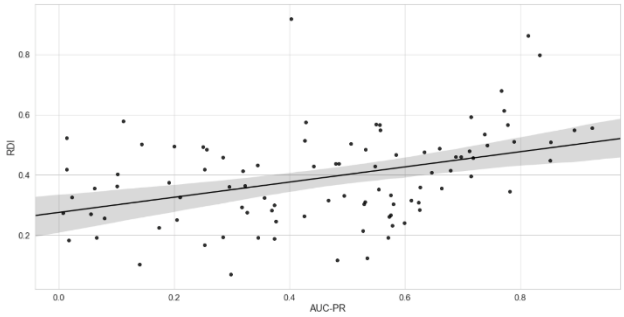


Figure 1: Subject RDI vs Model AUC-PR score.

4.3 Illustrative Examples

Figure 2 illustrates the model's detection of a RERA event from subject tr05-1042. The event manifests as a decrease in the amplitude of the Abdominal, Chest, and Airflow signals and its extent is established by solid black line labeled "RERA". The model's RERA probability rises before the onset of the event; peaks with the reduction in amplitude of the respiratory channels; and falls before the offset of the event.

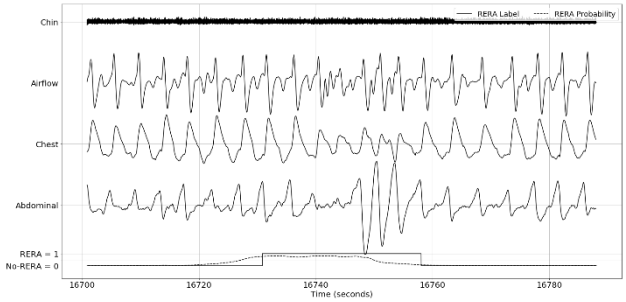


Figure 2: Singular RERA event detection.

Figure 3 below illustrates the model's detection of a sequence of RERA events from subject tr09-0082. The events again manifest as a decrease in the amplitude of the respiratory channels. Evidence of subject arousal can be seen in the Chin channel disturbance following the first RERA event. The model's RERA probability rises well above the 0.5 probability during the events and drops below that level between and after the events.

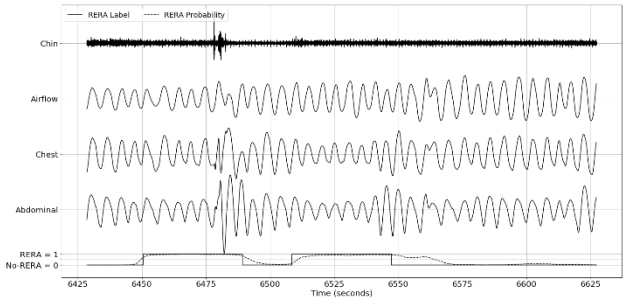


Figure 3: RERA sequence detection

Figure 4 shows a long RERA event from subject tr06-0103, with a repeating pattern of varying amplitudes in the respiratory channels. The model's RERA probability rises above the 0.5 probability level before the event and is sustained throughout the event.

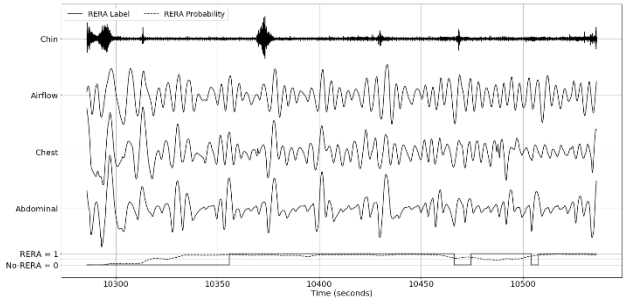


Figure 4: Long RERA detection

Figure 5 illustrates the model falsely detecting a RERA event from subject tr13-0387. In this example two RERA events (between 2625-2650 secs and 2675-2700 secs) are present and associated with respiratory waveform amplitude reductions. Then, breathing amplitude remains variable and reduced (2700-2750 secs). However, this is not associated with a RERA ground-truth label. The model correctly predicts the first two events but incorrectly rises again after the second RERA event.

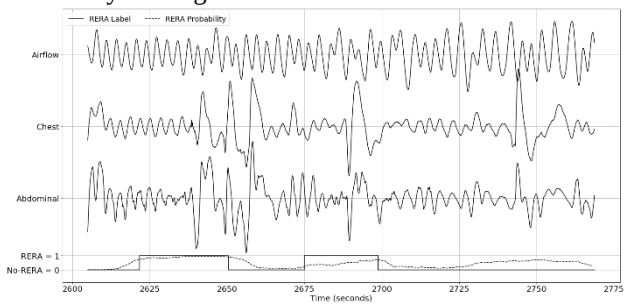


Figure 5: False RERA detection

Figure 6 is a t-SNE visualization of the activations of 1600 randomly sampled RERA and normal examples in the final layer before the classification layer. The orange dots are the RERA examples and blue are normal. While we see some separable clusters, a sizeable number of normal examples contaminate the RERA cluster decreasing the precision of RERA detection.

Figure 6: t-SNE visualization of final activation layer

5. Discussion and Conclusion

Unlike apneas and hypopneas, RERAs are subtle and variable in appearance as in Figures 2-5. This renders the task of automatically detecting RERAs a challenge.

The neural network evaluated successfully combined multi-modal feature time-series to detect singular, clustered, and long RERA events (Figure 2-4). However, the network had poor precision on subjects with a low burden of sleep-disordered breathing (Figure 1). The network mistook periods of reduced respiratory activity followed by a rebound as a RERA event (Figure 5).

Recall that a RERA is a sequence of respiratory events culminating in an arousal. This suggests that one should establish the presence of an arousal and then evaluate the respiratory disturbance. Our hypothesis is that the neural network considers evidence of an arousal as a correlate of a RERA rather than a requirement. Better encoding of the arousal contingency in the network could improve its precision to the level necessary for clinical deployment.

Even the model's strengths surface some interesting directions for future work. The wide features were not used suggesting that the RR intervals spectra and stages do not add more information to the model. Better experiment and explanatory techniques can be used to identify the most important and accurate features for detecting respiratory disturbance during sleep.

References

[1] Ferguson, Kathleen A; et al., "Consequences of Sleep Disordered Breathing," *Thorax*, vol. 50, pp. 998-1004, 1995.

[2] Ayappa, Indu; et al., "Non-Invasive Detection of Respiratory Effort-Related Arousals (RERAs) by Nasal Cannula/Pressure Transducer System," *Sleep*, vol. 23, no. 6, pp. 763-771, 2000.

[3] Berry, Robert B; et al., "Rules for Scoring Respiratory Events in Sleep: Update 2007 AASM Manual for the Scoring of Sleep and Associated Events," *Journal of Clinical Sleep Medicine*, vol. 8, no. 5, pp. 597-619, 2012.

[4] J.F, Masa; et al., "Assessment of Thoracoabdominal Bands to Detect Respiratory Effort-Related Arousal," *European Respiratory Journal*, vol. 22, pp. 661-667, 2003.

[5] Malhotra, Atul; et al., "Performance of Automated Polysomnography Scoring System Versus Computer-Assisted Manual Scoring," *Sleep*, vol. 36, no. 4, pp. 573-582, 2013.

[6] Selvaraj, Nandakumar; et al., "Detection of Sleep Apnea on a Per-Second Basis Using Respiratory Signals," in *35th Annual International Conference of the IEEE EMBS*, Osaka, Japan, 2013.

[7] Mohammad M Ghassemi; et al., "You Snooze, You Win: the PhysioNet/Computing in Cardiology Challenge 2018", *Computing in Cardiology Volume 45*. Maastricht, Netherlands, 2018. pp 1-4

Address for correspondence.

Niranjan Sridhar
Verily Life Sciences
269 East Grand Avenue, South San Francisco, CA, USA, 94080
nirsd@verily.com