# Prediction of Apnoea and Non-apnoea Arousals From the Polysomnogram Using a Neural Network Classifier

Philip de Chazal, John Du, Nadi Sadr

Charles Perkins Centre and School of Biomedical Engineering,
The University of Sydney, Australia

## Abstract

*In this study we present a system for automated processing of signals from the polysomnogram (PSG) for the detection of apnoea and non-apnoea arousals. The PSG signals were divided into 15 second epochs and 59 time- and frequency-domain features were derived for each epoch. Features from adjacent 4 epochs were combined and processed with a bank of ten feed-forward neural networks each with a single hidden layer of 20 units. The system outputs a 200 Hz annotation signal containing probability estimates that each sample was associated with an apnoea or non-apnoea arousal, or no-arousal. Data from the Physionet Computing in Cardiology Challenge 2018 was used to develop and test the system. Performance of the system was assessed using three class and two class metrics. With the system classifying three classes, the volume under the receiver operator characteristic (ROC) surface was 0.75 with an optimal specificity of 0.72, a sensitivity of 0.76 for the apnoea arousals, and a sensitivity of 0.69 for the non-apnoea arousals. When the two arousal classes were combined into one arousal class, the area under the precision recall curve was 0.74, the area under the ROC curve was 0.91, with an optimal specificity and sensitivity of 0.85.*

## 1. Introduction

A wide range of negative health outcomes including neurocognitive disorders, mood and mental conditions, cardiovascular disease [1], hypertension and stroke [2] are associated with low quality sleep. Poor sleep is also an established contributor to workplace and road accidents [3]. While arousals are a normal feature of the sleep/wake cycle, an excessive number of arousals can lead to poor sleep quality [4-6]. Respiration interruptions during sleep are a common cause of arousals. These interruptions include obstructive apnoea and hypopnoea events, respiratory effort related arousals and other interruptions to breathing. Arousals can also be caused by snoring, muscle jerks, pain, and insomnia.

The most common way to provide a detailed assessment of sleep is to record a polysomnogram (PSG) which provides range of signals from a sleeping patient [7]. Sleep technicians then manually assess the PSG. Part of their analysis includes scoring arousals which is a time-consuming manual task. Automated software that processes the PSG information and assists the technician in identifying arousals would clearly be of benefit and is the topic of this paper.

The PhysioNet Computing in Cardiology Challenge 2018 [8] provided the framework for researchers to develop and test automatic algorithms for the detection of non-apnoea arousals from the PSG and we were one of the participating teams [9,10]. For the purposes of the Challenge apnoea-related arousals were ignored which limited the clinical application of the resulting algorithms. In this current study we address this application issue and extend the capability of our system to detect apnoea arousals. Our resulting system potentially has greater clinical application.

## 2. Input data

The data used in this study dataset was provided by the 2018 Challenge organisers (https://physionet.org/challenge) and is publicly available. It includes 994 overnight PSG study recordings and associated sleep, respiratory event and arousal annotations [8]. All recordings were acquired at the Massachusetts General Hospital (MGH) sleep laboratories.

### 2.1. Signals and expert scorings

The signals include the airflow, electrocardiogram, electroencephalogram (EEG), chin electromyogram (EMG), electrooculogram (EOG), pulse oximetry (SpO2) and respiratory effort signals. All signals were sampled at 200Hz. Standard sleep and respiratory events were scored by MGH staff. They also determined the time, duration and cause of all arousal events. Arousals were then grouped into apnoea related arousals and non-apnoea related arousals. Apnoea arousals included central, mixed and obstructive apnoeas and hypopnoeas. Non-apnoea

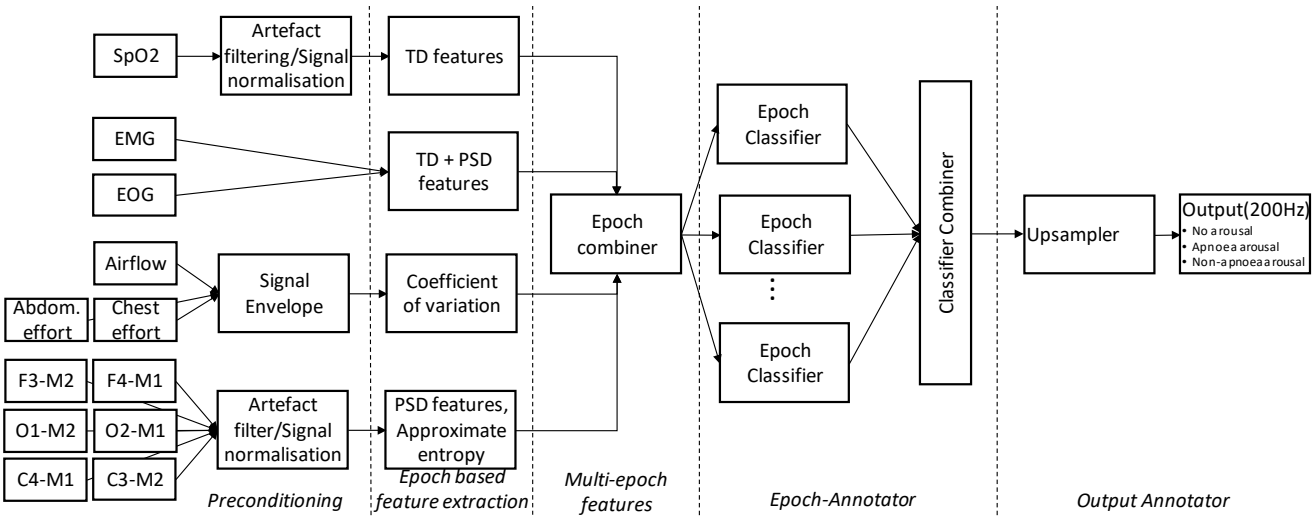


Figure 1. Outline of the proposed automatic arousal detection system annotating apnoea arousal, non-apnoea arousal and no-arousal events.

Table 1. Features extracted from the PSG signals.

| PSG signal(s) | Features | Number |
|---|---|---|
| SpO2 | Kurtosis. Hypoxic burden. Mean of absolute differences. Proportion: <90%, <92%, <94%. Skewness. Standard deviation. | 8 |
| Chin EMG<br>EOG | Form factor. Kurtosis. Relative band energy: 0-2Hz 2-4Hz. Skewness. Standard deviation. | 12 |
| Abdominal respiratory effort.<br>Airflow, Chest respiratory effort | Signal envelope coefficient of variation | 3 |
| EEG: F3-M2, F4-M1 O1-M2,<br>O2-M1, C4-M1, C3-M2 | Approximate entropy.<br>Relative band energy: 2-4Hz, 5-8Hz, 9-12Hz, 13-16Hz, 17-32Hz. | 36 |
| | **Total** | **59** |

related arousals included breathing, bruxisms, Cheyne-Stokes hypoventilations, partial airway obstructions, periodic leg movement, respiratory event related arousals, snores, spontaneous arousals, and vocalisations

Based on the arousal events, a target vector sampled at 200Hz was established for every PSG recording by the Challenge organisers. The target values for the non-apnoea arousals were set to "1". The target values of the apnoea (hypopnoea)-arousals were set to "-1". The target value for the non-arousals sections were set to "0" [8].

## 3. Methods

We adopted the same feature extraction framework as previously used by us in the 2018 Challenge. Features were extracted from 15 second time windows of data as previously we had shown that this provided the longest epoch length that maintained satisfactory precision of the arousal annotations. The epoch-based annotations needed for the training algorithms were determined from the 200Hz arousal annotations by sampling the arousal annotations at the midpoint of every 15 seconds epoch [9].

Figure 1 shows an outline of our proposed system. It processes SpO2, EMG chin, EOG, respiratory effort, airflow, and EEG signals from the overnight PSG recordings. The system first removes artefacts from signals and normalizes signals. Following this, the system divides the signals into 15 second epochs, calculate features per epoch, combines features from adjacent epochs, classifies each epoch using a bank of classifiers and then finds a combined classifier output for each epoch. The final step in the process is to upsample the epoch annotations to a 200Hz output signal. The annotation per sample is one of "no arousal", "apnoea arousal" or "non-apnoea arousal".

### 3.1. Features and epoch combiner

The features extracted per epoch from the PSG are shown in Table 1. Full details of the feature extraction methods are provided in [9].

The system was provided with the ability to look forward/backwards in time by up to one minute for signal changes associated with the arousals. This was achieved by combining features from four epochs either side of the

Table 2. Performance results* the LDA and FFNN classifier. Results are shown for the three class and two class configurations. The two class results are obtained by merging the non-apnoea arousal and apnoea arousal results into one arousal class.

| | Three classes (Non-arousal, Non-apnoea arousal, Apnoea-arousal) | | | | | | Two classes (Non-arousal, Arousal) | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | AUPRC | | | | | |
| Classifier | VUROS | Non-arousal spec. | Non-apnoea arousal sens. | Apnoea-arousal sens. | Non- apnoea arousal | Apnoea-arousal | AUROC | Non-arousal spec. | Arousal sens. | AUPRC Arousal |
| LDA | 0.68 | 0.68 | 0.71 | 0.65 | 0.14 | 0.71 | 0.88 | 0.81 | 0.82 | 0.67 |
| FFNN | 0.75 | 0.72 | 0.69 | 0.76 | 0.18 | 0.77 | 0.91 | 0.85 | 0.85 | 0.74 |

Table 3. Normalised confusion matrix* for three classes at the optimum specificity and sensitivities using the FFNN classifier.

| | | Predicted class | | |
|---|---|---|---|---|
| | | Non-arousal | Non-apnoea arousal | Apnoea arousal |
| Expert Class | Non-arousal | 0.72 | 0.22 | 0.06 |
| | Non-apnoea arousal | 0.17 | 0.69 | 0.14 |
| | Apnoea arousal | 0.05 | 0.18 | 0.77 |

Table 4: Normalised confusion matrix* for two classes at the optimum specificity and sensitivity using the FFNN classifier.

| | | Predicted class | |
|---|---|---|---|
| | | Non-arousal | Arousal |
| Expert Class | Non-arousal | 0.85 | 0.15 |
| | Arousal | 0.15 | 0.85 |

Abbreviations: AUPRC: Area Under the Precision-Recall Curve, AUROC: Area under the ROC curve. FFNN: Feedforward neural network; LDA: Linear discriminant analysis; Sens: Sensitivity; Spec: Specificity; VUROS: Volume under the ROC surface.

*Calculated from test-set results of ten-fold cross validation.

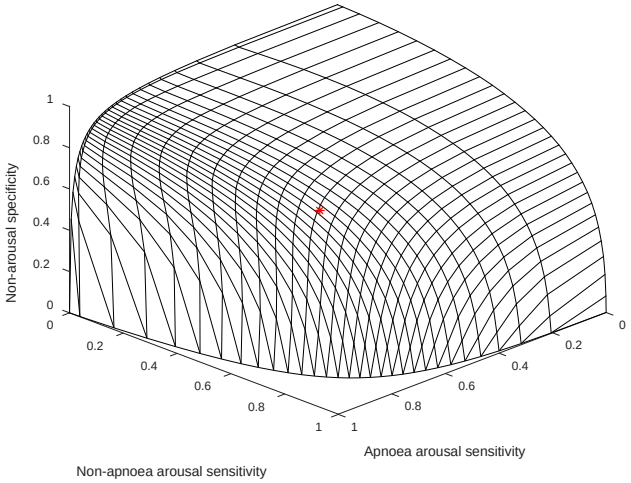


Figure 2: Three-way receiver operator characteristic surface (ROS) for the test set results of the FFNN system. The volume under the ROS is 0.75. The optimum specificity/sensitivity points are indicated with the red marker.

current epoch which were then processing by the classifier stage. Four epochs were c hosen as previous work by us [11] had shown this provided good performance.

## 3.3. Classifiers and upsampler

Two classifiers were used to provide a performance comparison. A linear discriminant analysis (LDA) classifier [12], and bank of 10 single hidden layer feed-forward neural networks (FFNN) [13] each with 20 hidden units. Both classifiers had a softmax output stage that provided three probability outputs of non-arousal, apnoea arousal and non-aponea arousal for every epoch. The LDA classifier was fast to train, while the FFNN classifier had provided superior performance in the 2018 Physionet Challenge [9]. The FFNN was trained with a cross-entropy cost function.

The outputs of the bank of FFNN classifiers were combined by averaging the probability estimates of the ten individual classifiers.

The final stage was to upsample the epoch classifications to the 200Hz sample rate of the event based annotations using a first order hold filter.

## 3.5. Performance estimation

Performance was estimated using 10-fold cross-validation. The 994 records were randomly divided into ten sets containing 99 or 100 records. Nine sets were used to train the system and the remaining set used to test performance. The sets were then rotated and the combined performance results on the test sets calculated.

Performance was measured using several metrics. First, 3×3 confusion matrices were calculated over the complete range of decision thresholds for the three classes. The specificity and sensitivity results from the confusion matrices were then used as inputs to a three-way receiver operator characteristic (ROC) analysis [14]. A summary performance measure was calculated by finding the volume under ROC surface (VUROS) which is the 3 class equivalent of the area under the ROC curve (AUROC).

The three-way ROC was also used to find optimal specificity and sensitivity points by finding the point on the surface that minimized the Euclidean distance to the ideal performance point of (1 1 1). The sensitivity and positive predictive results from the 3×3 confusion matrices were used to perform a Precision-Recall curve

analysis of the two arousal classes [15]. Performance was assessed using area under the curves (AUPRC).

Second, the cell entries of the two arousal classes in the 3×3 confusion matrices were combined, producing a 2×2 confusion matrices representing the performance at each decision threshold for a system classifying “non-arousal” and “arousal”. The AUROC, AUPRC and optimal specificity and sensitivity points were calculated from the 2×2 confusion matrices.

## 4. Results and discussion

Table 2 shows the key results for the two classifiers. The performance of the FFNN outperformed the LDA in all measures. For the three class results, the FFNN classifier achieved a VUROS of 0.75, a specificity of 72%, an apnoea arousal sensitivity of 76% and a non-apnoea arousal of 69%. The AUPRC was 0.18 and 0.76 for the non-apnoea arousal and apnoea-arousal class respectively. Our results suggest that our algorithm detects apnoea-arousals more effectively than non-apnoea arousals. Possible reasons are that apnoea arousals occurred more frequently in the database than the non-apnoea arousals. Also, non-apnoea arousals have diverse causes which results in greater heterogeneity of signals.

Figure 2 shows the test-set 3-way ROC surface with the optimum performance point indicated. Table 3 shows the normalised confusion matrix for the three classes at the optimum point determined from the ROC surface. It shows that non-arousal and the apnoea arousal classes tended to be misclassified as the non-apnoea arousal class with false detection rates of 0.22 and 0.18 respectively. After combining the two arousal classes into one class, the AUROC was 0.91 with an optimal specificity/sensitivity of 0.85, and an AUPRC of 0.74 (see Tables 2 and 4).

There are several areas for future work. Our system used 531 features which very likely includes some highly correlated features, so feature selection may yield a smaller higher performing system. The highest performing architectures from the 2018 Physionet Challenge included convolutional neural network stages and used minimally processed sensor data. Adopting similar architectures may yield performance gains.

## 5. Discussion and Conclusion

We’ve presented a system for automated annotation of selected signals from the PSG for the presence of apnoea and non-aponea arousals. Our best system used a bank of 10 feed-forward neural networks and achieved a volume under the ROC surface of 0.75 with a specificity of 72%, a sensitivity of 76% for the apnoea arousals, and a sensitivity of 69% for the non-apnoea arousals. When the two arousal classes were combined into one it achieved an AUROC of 0.91 and an AUPRC of 0.74.

## References


[1] Hamilton GS, Solin P, Naughton MT. Obstructive sleep apnoea and cardiovascular disease. Internal medicine journal. 2004;34(7):420-6.

[2] Young T, Peppard PE, Gottlieb DJ. Epidemiology of obstructive sleep apnea: a population health perspective. American journal of respiratory and critical care medicine. 2002;165(9):1217-39.

[3] Leger D. The cost of sleep-related accidents: A report for the National Commission on Sleep Disorders Research Sleep 1994;17(1):84–93.

[4] Mathur R, Douglas NJ. Frequency of EEG arousals from nocturnal sleep in normal subjects. Sleep 1995;18(5):330–333.

[5] Dahl RE. The regulation of sleep and arousal: Development and psychopathology. Development and psychopathology. 1996;8(1):3-27.

[6] De Carli F, Nobili L, Gelcich P, Ferrillo F. A method for the automatic detection of arousals during sleep. Sleep. 1999;22(5):561-72.

[7] Lim DC, Mazzotti DR, Sutherland K, Mindel JW, Kim J, Cistulli PA, Magalang UJ, Pack AI, de Chazal P, Penzel T. Reinventing Polysomnography in the Age of Precision Medicine. Sleep Medicine Reviews. 2020:101313.

[8] Ghassemi MM, Moody BE, Lehman LW, Song C, Li Q, Sun H, Mark RG, Westover MB, Clifford GD. You snooze, you win: the physionet/computing in cardiology challenge 2018. Computing in Cardiology 2018;45:1-4.

[9] de Chazal P, Sadr N. Automatic scoring of non-apnoea arousals using hand-crafted features from the polysomnogram. Physiological Measurement. 2019;40(12):124001.

[10] Sadr N, de Chazal P. Automatic scoring of non-apnoea arousals using the polysomnogram. Computing in Cardiology 2018;45:1-4.

[11] de Chazal P, Sadr N. Automated Annotation of Polysomnogram Epochs for Apnoea and Non-apnoea Arousals. 42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) 2020; 2796-2799.

[12] de Chazal P, Heneghan C, Sheridan E, Reilly R, Nolan P, O'Malley M. Automated processing of the single-lead electrocardiogram for the detection of obstructive sleep apnoea. IEEE Transactions on Biomedical Engineering. 2003 Jun 11;50(6):686-96.

[13] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media; 2009.

[14] Mossman D. Three-way rocs. Medical Decision Making. 1999;19(1):78-89.

[15] Boyd K, Eng KH, Page CD. Area under the precision-recall curve: point estimates and confidence intervals. Joint European conference on machine learning and knowledge discovery in databases 2013:451-466.



Address for correspondence.

Philip de Chazal.
Charles Perkin Centre and School of Biomedical Engineering
The University of Sydney, NSW 2006, Australia.
philip.dechazal@sydney.edu.au.