Efficacy & Implementation of Automatic Cortical Arousal Detector by Brink-Kjaer et al. 2020 to ANPHY Data

Joon Hwan Hong, 260832806

Introduction

Electrophysiological activity in the brain can be detected using an electroencephalogram (EEG). Electrodes attached to the scalp detect changes in brain activity with high temporal resolution. They measure the change in voltage over time from the ionic currents generated from neurons. An EEG signal is decomposed of five frequency bands: delta (0.5 to 4Hz), theta (4 to 7Hz), alpha (8 to 12Hz), beta (13 to 30Hz), and gamma (30 to 80Hz) to classify changes in brain states (Nayak & Anilkumar, 2020). Consequently, EEG recordings are able to detect changes in wakefulness in the brain based on their activity patterns.

Sleep is a heterogeneous physiological process: it oscillates between rapid eye movement (REM) sleep and non-rapid eye movement (NREM) sleep stages (Carley & Farabi, 2016). The study of modern sleep science and the discovery of sleep macro- & micro-architectures are closely connected to the history of EEG development. Distinct EEG patterns constituting NREM sleep – vertex waves, sleep spindles, K complexes, etc. – were discovered by Loomis and others in 1937 characterizing EEG patterns (Shepard et al., 2005). The development of EEG allowed for the identification of an electrophysiologic substrate of sleep. Today, multimodal integration of EEG and other signals derived from the human body allows for further study of sleep and brain activity.

Background

The standard for assessing sleep and wake states is done by polysomnography (PSG), a type of sleep study where brain waves, oxygen level in blood, heart rate, breathing, as well as leg and eye movements are recorded during sleep (Keenan, 2005). In PSG recordings, an electrophysiological phenomenon called arousals can be detected. They are abrupt changes detected in EEG, showing brief intrusions of wakefulness during sleep; they present states of transitions, providing a unique window to understand the sleep-wake boundary regulation (Brink-Kjaer et al., 2020). A case where this would be valuable is dysregulation of sleep cycles resulting in severe daytime sleepiness in some patients while not in others. Therefore, arousals provide a means to investigate the dysregulation of sleep for patients with sleep disorders. An explanation could be that sleep is regulated at a local level, resulting in distinct participation of various brain structures in arousal generation (Nir et al., 2011; von Ellenrieder et al., 2020).

To study sleep microstructures like arousals, clinical sleep scoring requires a visual review of PSG recordings by a human expert. However, this results in high inter-scorer variability and low intra-scorer agreement due to varied interpretations of sleep scoring guidelines (Fiorillo et al., 2019). Hence a fully automatic, robust, standardized method for detecting arousals is a method to be sought after. Recently,

a deep learning tool for automatic analysis of sleep scoring and arousal detection – developed by the lab of Emmanuel Mignot and others – was published for PSG recordings (Brink-Kjaer et al., 2020). While functional arousal detection algorithms have been developed previously, many systems are validated either using small datasets (6-60 PSGs), or had few human scorers resulting in the lack if effective out-of-distribution generalization (Brink-Kjaer et al., 2020). The overarching goal of this project is to observe the efficacy of automated arousal detection, seeing if it is in agreement with data generated in the lab. The number of arousals detected and duration – any cases of disagreement and why – are particular points of interest.

Methods

The ANPHY database is used as the project dataset. It is comprised of 45 recordings which 11 files meet the criteria of all needed signals except left and right separated EOG signals for the detector. Instead, the files have a single EOG signal which is used instead for both. Data is stored as *.edf* (European Data Format) filetypes. EEG can be monitored either by bipolar montage or referential montage. Depending on the reference, amplitude can differ. A referential montage would indicate a common reference electrode for all the channels recorded; a bipolar montage would indicate that two electrodes are used per channel, referenced to another electrode. For the project, bipolar sleep montages are used. To allow for consistent quality of physiological signals, the following preprocessing steps are followed: loading the edf, resampling the data, filtering the data for artifacts, then exporting the results as a text file.

The system requires a central EEG derivation (C3 or A2), left and right EOG (electrooculogram), chin EMG (electromyogram), and ECG (electrocardiogram) (Brink-Kjaer et al., 2020). For the project, various definitions of a central EEG signal are tested: C3-A2, C3-M2, and C3. The *PreprocessNewData.m* script calls *LoadEDF.m* as a child process, which uses the publicly available *edfRead* script. Within the *LoadEDF* script the central EEG, left EOG, and right EOG signals are derived. However, for the project only one derivation for EOG is available. This is due to the fact that only two electrodes were used – one at the top of the left eye and the other on the bottom of the right eye – resulting in only one derivation instead of two to separate left and right EOG signals. Consequently, the singular EOG signal is fed twice as a replacement. The second function of *LoadEDF* is to detect if any of the signals are not recorded in µV scaling, and convert if necessary.

The output from *LoadEDF* is then directed to preprocessing. The process is composed of two sequential scripts*, preprocess.resampledata* and *preprocess.filterData*. All signals are resampled to 128Hz. To avoid aliasing effects, the misidentification of signal frequency – a low pass least-squares Finite Impulse Response (FIR) filter is applied (Brink-Kjaer et al., 2020). The FIR filter minimizes the weighted squared error between the filter's magnitude response and the idea piecewise function over the frequency bands. Subsequently, the processed header and data objects are passed through a bandpass filter in *preprocess.filterData*. Infinite Impulse Response (IIR) filters are used to remove frequency content that represent powerline interference in EEG and EOG. A bandpass filter allows

frequencies within a desired/selected range and rejects values outside the range. A passband range of 0.5-35Hz is chosen to preserve most physiologically meaningful data while removing frequencies from potential power line interferences and low frequency artifacts (Brink-Kjaer et al., 2020). A previously implemented Recursive Least Square (RLS) adaptive filter is used to remove ECG and ocular movement artifacts in EEG (Iv et al., 2014). Finally, *exportData.m* script is used to export the final data as a text file.

The arousal detector is a convolutional and LTSM (long short-term memory) recurrent neural network. The neural network functions by taking a static text output from *exportData* and processing it with a network of filters. LTSM recurrent networks are selected as they are capable of modeling long-term dependencies while avoiding the issues of exploding and vanishing gradients (Hochreiter & Schmidhuber, 1997). The overall network architecture uses the convolutional neural network to automatically generate features in select timestep bins, and subsequently fed to LTSM network followed by a fully connected neural network dedicated for predictions of arousal and wake as probabilities (Brink-Kjaer et al., 2020). The trained network architecture is provided by the paper. The model is accessed by defining string flags for the model settings in *ar_predict.py* and running the script.

The prediction output is formatted as a [1x*n*] array, where *n* is the length of the recording in seconds (128 samples). Each cell contains a predicted probability that the 1 second bin contains an arousal; the probability is binarized by a probability threshold. Various binarization thresholds are tested: 0.05, 0.10, and 0.15. The arousal detector does not distinguish wake from sleep, and results in false arousal detections in wake stages. *removeWake()* function then removes arousal detections using SigFI scripts. Detections that overlap into wake stage remains. To observe if predicted arousals are in agreement with scored arousals, *closeness(ϑ)* function checks given a boundary θ if ∃ predicted arousal such that it is in the range (arousal ± θ).

Overall, for the project: *LoadEDF* script, *preprocess.resampleData*, and *exportData* are modified for compatibility with the lab's PSG data; *reformatting(), removeWake(), extractTimes(), closeness(), multi_add_event()* functions are created to analyze and upload the results into .STS files (data format for Stellate Harmonie software) to review visually in the Stellate Reviewer software. As the stellate reviewer-MATLAB I/O is only compatible with MATLAB 2010b, additional scripts are implemented to call I/O commands from MATLAB 2021a to a MATLAB 2010b child process in *run_2010.m*. The preprocessing and postprocessing pipeline are reorganized as *Preprocessing.m* and *Postprocessing.m* for ease of future use. Project Repository is available at: https://github.com/Joon-Hwan-Hong/Arousal-Detector-ANPHY-Integration.

Results

The number of predictions within five seconds of a scored arousal is used as a proxy to the detector's prediction efficacy. Preliminary results from the recommended binarization threshold of 0.225 in the literature results in under-detection. Lower thresholds are tested. Results from varying recordings and binarization thresholds ranging 0.05 to 0.15 are shown in Table 1. While decreasing the threshold alleviated the issue of under-detection in quantity, most predicted arousals are not in

agreement with marked arousals. In the FA536671 recording, only with CA-A2 central EEG definition with binarization threshold of 0.05 results in any arousal prediction within 5 seconds of a marked arousal timestep.

Name	CEEG	Threshold	# arousals	# predicted	Difference	# pred in ±5s	# pred in ±10s	Arousals Missed %	Arousals Detected %
FA536671	C3	0.05	26	22	4	0	0	100.00	0.00
	C3	0.08		17	9	0	0	100.00	0.00
	C3	0.15		12	14	0	0	100.00	0.00
	C3-M2	0.05		30	-4	0	0	100.00	0.00
	C3-M2	0.10		23	3	0	0	100.00	0.00
	C3-M2	0.15		17	9	0	0	100.00	0.00
	C3-A2	0.05		7	19	1	1	96.15	3.85
EP1373_1BF	C3-A2	0.05	28	26	2	3	7	75.00	25.00
	C3-A2	0.10		19	9	2	5	82.14	17.86
	C3-A2	0.15		13	15	2	3	89.29	10.71
EP1373_2BF	C3-A2	0.05	59	70	-11	9	15	74.58	25.42
	C3-A2	0.10		46	13	8	12	79.66	20.34
	C3-A2	0.15		32	27	7	12	79.66	20.34

Table 1: Raw Results of the Multimodal Arousal Detector when fed with preprocessed data from three recordings (FA536671, EP1373_1BF, EP1373_2BF). Various definitions of central EEG are tested for FA53667; C3-M2 and C3-A2 are bipolar recordings. Thresholds are the binarization probability values chosen. 'Difference' column is the magnitude difference between the number of marked arousals and predicted. '# pred in ±5s' is the number of predicted arousals within 5 seconds of a marked arousal.

Name	CEEG	Threshold	TP	FP	FN	Precision	Recall	F1 Score
FA536671	C3	0.05	0	22	26	0	0	0
	C3	0.08	0	17	26	0	0	0
	C3	0.15	0	12	26	0	0	0
	C3-M2	0.05	0	30	26	0	0	0
	C3-M2	0.10	0	23	26	0	0	0
	C3-M2	0.15	0	17	26	0	0	0
	C3-A2	0.05	1	6	25	0.143	0.038	0.061
EP1373_1BF	C3-A2	0.05	3	23	25	0.115	0.107	0.111
	C3-A2	0.10	2	17	26	0.105	0.071	0.085
	C3-A2	0.15	2	11	26	0.154	0.071	0.098
EP1373_2BF	C3-A2	0.05	9	61	50	0.129	0.153	0.140
	C3-A2	0.10	8	38	51	0.174	0.136	0.152
	C3-A2	0.15	7	25	52	0.219	0.119	0.154

Table 2: Precision, recall, and F1 score of the multimodal arousal detection. TP (True Positive) in this context is defined to be the '# of pred in ±5s' (column in *Table 1*). FP (False Positive) is defined to be the difference between the number of predicted arousals and '# of pred in ±5s'. FN (False Negative) is defined to be the difference between the number of marked arousals and '# of pred in ±5s'. F1 score is the harmonic mean of precision and recall.

Inefficacy is similarly observed for other recordings regardless of the tested thresholds. In EP1373_1BF recording, there are 28 scored arousals. Threshold of 0.05 results in three predicted arousals in close proximity to marked arousals. While the other two thresholds tested – 0.10 and 0.15 – results in two. In EP1373_2BF recording, there are 59 scored arousals. The trend continues as only 9, 8, and 7 (for 0.05, 0.10, and 0.15 respectively) predictions are within the five second window from marked arousals.

Low precision, recall, and F1 score are observed for all trials, seen in Table 2. The highest precision achieved is 0.219 from EP1373_2BF recording with a threshold of 0.15. The highest recall is 0.153 from the EP1373_2BF recording with a 0.05 threshold. The highest F1 score is achieved with a threshold of 0.15. However, the calculated precision, recall, and F1 score are determined on the assumption that all arousals are marked. Upon visual inspection of detections with Dr. Frauscher, it appears that many of the model predictions can be considered arousals which the previous marker missed. Two examples are shown in Figure 1. The PSG recordings are reviewed on the Stellate Reviewer software.

Discussion

Evidently, when given preprocessed PSG recordings from the Montreal Neurological Institute (MNI), the arousal detector does not perform to a satisfactory level; this is true if the definition of success is considered to be the number of predictions in agreement with marked arousals. Most arousal predictions are not in agreement with marked arousals in all recording samples tested. The results are obtained with low binarization thresholds; when a threshold of 0.05 is selected, all 1 second bins where the detector believed there is a 5% chance or higher of an arousal is set to be an arousal. Although to note, that the original machine learning architecture used a threshold of 0.225 (22.5%) which is the value which maximized the F1 score in their training set (Brink-Kjaer et al., 2020). There are potential issues of the project's implementation of the multimodal arousal detector which would contribute to the lack of agreement with scored arousals.

A distributional shift in PSG data – and expected input quality – can be a major obstacle when transferring machine learning prediction systems to unseen data. However, data diversity is sufficient in the original model. It is ensured by using thousands of subjects from various sleep cohorts. The original model is trained on 2889 MrOS and Cleveland Family Study PSGs, and the testing dataset included hundreds of unseen data from the Wisconsin (WSC) and Stanford Sleep Cohort (WSC). The decrease in performance on unseen data is reported to be acceptable (0.72 to 0.62 in precision, 0.76 to 0.7 in F1 score). Consequently, out-of-distribution generalization of the original model is shown to be effective to an extent. However, this is dependent on what channel signals are given to the preprocessing pipeline as the five input signals. Depending on the institution the data is from, varying definitions of input signals are selected in LoadEDF – such as varying EEG or EMG bipolar montage signals. Further optimization in the choice of input signals could improve detector performance. However, this is beyond the timeframe of the COMP 401 duration: a greater number of recordings beyond the current trials needs to be tested.

Another issue which is encountered in the project is the number of samples currently available to use. The available channels and their labels are different depending on the original purpose of recording in the lab. Not all provided recordings at the MNI have the minimum signals required for the arousal detector. Notably, many are lacking an ECG/EKG signal. Accordingly, retraining the model with the data provided is not plausible given the course timeframe due to: low number of PSGs immediately available and variable channel labeling depending on previous purpose of recording.

PSG recordings that are available for the project uses a single EOG recording. The model expects the EOG inputs to be separated into left and right eye EOG signals. While it is possible that inputting a single EOG signal for both can result in the mispredictions, left and right eye EOG signals are not independent from one another.

Seen in Figure 1, some predictions that would be represented as false positives can actually be considered as arousals (given the definition of true positive is the agreement of scored and predicted). It appears that some of the "mispredictions" are correct arousal detections not scored previously. Entirely relying on scored-predicted agreement as a measure of performance is inadequate given that not all arousals are scored previously. It is suggested that more Stereoelectroencephalography (SEEG) recordings are ran on the automatic arousal detector, and visually inspect the detected arousals to get an improved idea of the performance of the detector.

Conclusion

Implementation of the arousal detector to a new dataset is challenging: not all data samples meet the minimum input data requirements, different sampling frequencies, and other features can vary from the hardware used to obtain electrophysiological data.

While many detections are not in agreement with previously marked arousals in the three tested recordings, upon visual inspection promise has been shown in detections that can be agreed as an arousal. While further validation is needed, it seems at least in some files the detector is functional in initial visual examination. Future adjustments – choice of EEG input, binarization threshold, etc. – should be made to optimize the detector for the ANPHY dataset.

Figures

Figure 1: Example Visual Inspection of Predicted Arousals (Sample: FA536671, Threshold=0.05)

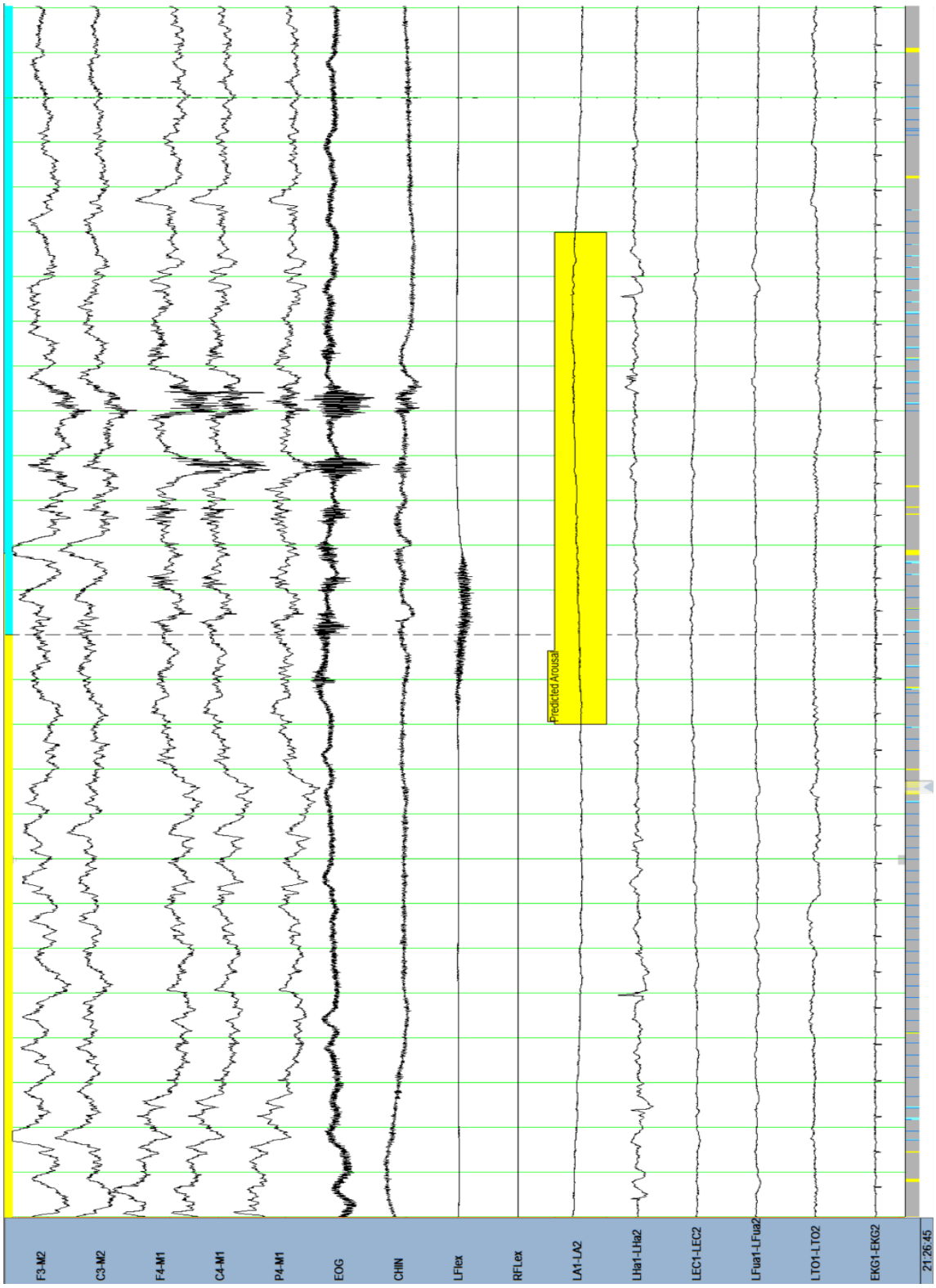
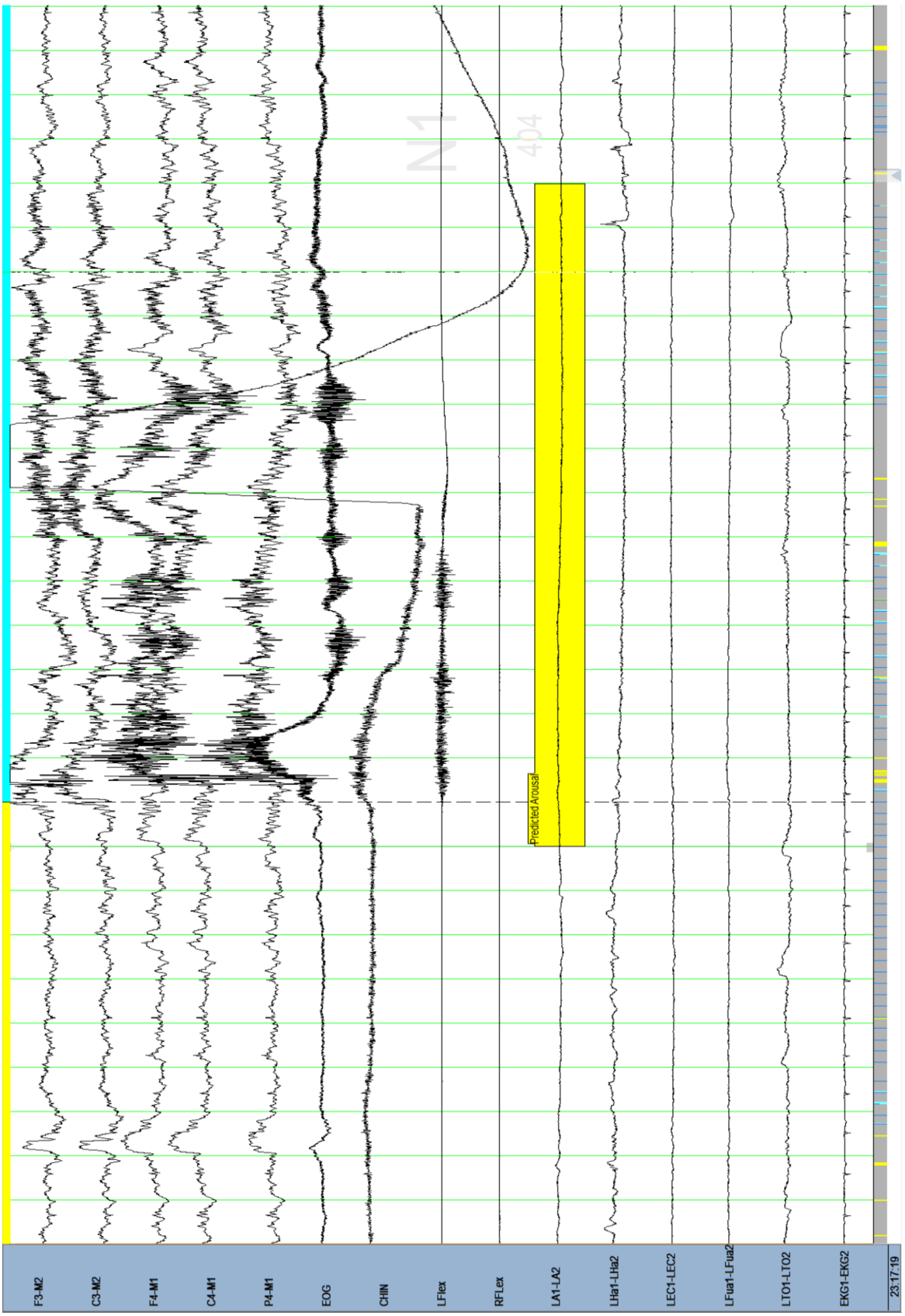
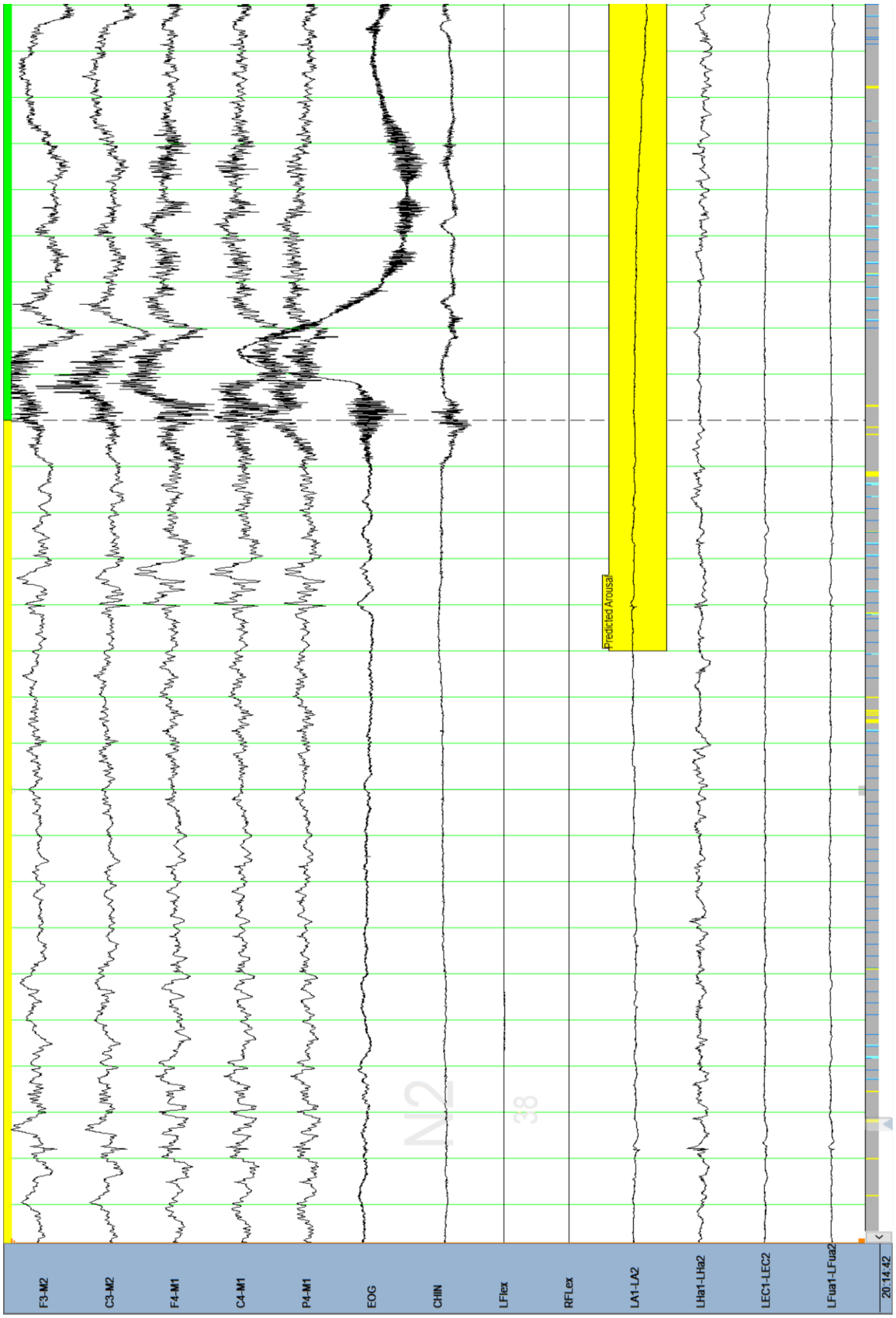


Figure 1: While not in agreement with a marked arousal, the predicted interval for an arousal fits the description. The arousal can be visually inspected such that an abrupt change in brain wave pattern is observed in bipolar EEG signals (F3-M2, C3-M2, etc.) & associated with brief increases in muscle tone in EMG signals (CHIN, LFlex, RFlex), which is an important feature for scoring arousals.


F3-M2
C3-M2
F4-M1
C4-M1
P4-M1
EOG
CHIN
LFlex
RFLex
LA1-LA2
LHa1-LHa2
LEC1-LEC2
LFua1-LFua2
LTO1-LTO2
EKG1-EKG2
Predicted Arousal
N1
404
23:17:19


F3-M2
C3-M2
F4-M1
C4-M1
P4-M1
EOG
ChIN
LFlex
RFLex
LA1-LA2
LHa1-LHa2
LEC1-LEC2
LFua1-LFua2
20:14:42
Predicted Arousal
N2
38

References

Brink-Kjaer, A., Olesen, A. N., Peppard, P. E., Stone, K. L., Jennum, P., Mignot, E., & Sorensen, H. B. D. (2020). Automatic detection of cortical arousals in sleep and their contribution to daytime sleepiness. *Clinical Neurophysiology*, *131*(6), 1187–1203. https://doi.org/10.1016/j.clinph.2020.02.027

Carley, D. W., & Farabi, S. S. (2016). Physiology of Sleep. *Diabetes Spectrum*, *29*(1). https://doi.org/10.2337/diaspect.29.1.5

Fiorillo, L., Puiatti, A., Papandrea, M., Ratti, P. L., Favaro, P., Roth, C., Bargiotas, P., Bassetti, C. L., & Faraci, F. D. (2019). Automated sleep scoring: A review of the latest approaches. In *Sleep Medicine Reviews* (Vol. 48, p. 101204). W.B. Saunders Ltd. https://doi.org/10.1016/j.smrv.2019.07.007

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Iv, H. M., Leary, E., Lee, S. Y., Carrillo, O., Stubbs, R., Peppard, P., Young, T., Widrow, B., & Mignot, E. (2014). Design and validation of a periodic leg movement detector. *PLoS ONE*, *9*(12), e114565. https://doi.org/10.1371/journal.pone.0114565

Keenan, S. A. (2005). Chapter 3 An overview of polysomnography. *Handbook of Clinical Neurophysiology*, *6*(C), 33–50. https://doi.org/10.1016/S1567-4231(09)70028-0

Nayak, C. S., & Anilkumar, A. C. (2020). EEG normal waveforms. In: StatPearls. In *StatPearls* (pp. 1–6). StatPearls Publishing. http://www.ncbi.nlm.nih.gov/pubmed/30969627

Nir, Y., Staba, R. J., Andrillon, T., Vyazovskiy, V. V., Cirelli, C., Fried, I., & Tononi, G. (2011). Regional Slow Waves and Spindles in Human Sleep. *Neuron*, *70*(1), 153–169. https://doi.org/10.1016/j.neuron.2011.02.043

Shepard, J. W., Buysse, D. J., Chesson, A. L., Dement, W. C., Goldberg, R., Guilleminault, C., Harris, C. D., Iber, C., Mignot, E., Mitler, M. M., Moore, K. E., Phillips, B. A., Quan, S. F., Rosenberg, R. S., Roth, T., Schmidt, H. S., Silber, M. H., Walsh, J. K., & White, D. P. (2005). History of the development of sleep medicine in the United States. *Journal of Clinical Sleep Medicine : JCSM : Official Publication of the American Academy of Sleep Medicine*, *1*(1), 61–82. https://doi.org/10.5664/jcsm.26298

von Ellenrieder, N., Gotman, J., Zelmann, R., Rogers, C., Nguyen, D. K., Kahane, P., Dubeau, F., & Frauscher, B. (2020). How the Human Brain Sleeps: Direct Cortical Recordings of Normal Brain Activity. *Annals of Neurology*, *87*(2), 289–301. https://doi.org/10.1002/ana.25651