Real-time Smartphone-based Sleep Staging using 1-Channel EEG

Abhay Koushik
MIT Media Lab, Cambridge USA
koushika@mit.edu

Judith Amores
MIT Media Lab, Cambridge USA
amores@media.mit.edu

Pattie Maes
MIT Media Lab, Cambridge USA
pattie@media.mit.edu

Abstract—**Automatic and real-time sleep scoring is necessary to develop user interfaces that trigger stimuli in specific sleep stages. However, most automatic sleep scoring systems have been focused on offline data analysis. We present the first, real-time sleep staging system that uses deep learning without the need for servers in a smartphone application for a wearable EEG. We employ real-time adaptation of a single channel Electroencephalography (EEG) to infer from a Time-Distributed Convolutional Neural Network (CNN). Polysomnography (PSG) —the gold standard for sleep staging—requires a human scorer and is both complex and resource-intensive. Our work demonstrates an end-to-end, smartphone-based pipeline that can infer sleep stages in just single 30-second epochs, with an overall accuracy of 83.5% on 20-fold cross validation for 5-stage classification of sleep stages using the open Sleep-EDF dataset. For comparison, inter-rater reliability among sleep-scoring experts is about 80% (Cohen's κ = 0.68 to 0.76). We further propose an on-device metric independent of the deep learning model which increases the average accuracy of classifying deep-sleep (N3) to more than 97.2% on 4 test nights using power spectral analysis.**

I. INTRODUCTION

Identification of sleep stages is important not only in diagnosing and treating sleep disorders but also for understanding the neuroscience of healthy sleep. PSG is used in hospitals to study sleep and diagnose sleep disorders and involves the recording of multiple electrophysiological signals from the body, such as brain activity using EEG, heart rhythm through Electrocardiography (ECG), muscle tone through Electromyography (EMG) and eye-movement through Electrooculography (EOG). PSG involves a tedious procedure which requires skilled sleep technologists in a laboratory setting. It requires a minimum of 22 wires attached to the body in order to monitor sleep activity. The complexity of this setup requires sleeping in a hospital or laboratory with an expert monitoring and scoring signals in real-time. This results in an unnatural sleep setup for the subject that affects the diagnosis and causes sub-optimal utilization of time and energy resources for recording and scoring. Therefore, there has been a significant development in research on automating sleep staging with wireless signals [31] and more compact, wearable devices [7], [20], [22]. Nevertheless, as far as we are aware of, none of these systems implements a 5-stage classification of sleep in just 30-second real-time epochs on a smartphone using single-channel EEG.

The goal of our research was to simplify and automate PSG on a smartphone for automatic and real-time interventions which can potentially be used in future Human-Computer-Interaction (HCI) applications. Automated classification is achieved through adaptation of a Time-Distributed Deep Convolutional Neural Network model to classify the 5 sleep stages. As per the new AASM rules [4], these stages are—Wake, Rapid-Eye-Movement (REM) and Non-Rapid-Eye-Movement (N-REM) stages N1, N2, and N3. We make use of a single channel EEG recorded through a modified research version of the Muse headband [10]. We have developed a TensorFlow Lite Android app that uses only a single channel recording of EEG. The app has a friendly user interface to visualize sleep stages and EEG data with real-time statistics. It connects via Bluetooth Low Energy (BLE) to the flexible EEG headband making it portable and comfortable to use at home.

II. RELATED WORK

Automatic identification of sleep stages through EEG feature extractions have been explored in the past [23]. More recent work has focused on simplifying sleep staging systems into portable forms that use motion sensing, heart-rate, pulse-oximetry and respiration levels [7], [22]. Zhang et al. [30], uses pulse, blood oxygen and motion sensors to predict sleep stages. The authors mention that they do not detect sleep stages N1 and N2 separately and that these results cannot provide equally high accuracy as compared to the EEG and EOG signals of PSG. The same limitations apply to the work by Zhao et al [31]. Previous researchers have looked into detecting sleep stages using different types of EEG form factors, such as stickers or in-ear EEG [26] [14] [15]. Our work differs from these works by being the first that computes 5-stage classification using one lead, on a 30-second epoch and processes the data in real-time on a smartphone without the need for server-client architecture as used in Dreem headband [18]. Our TensorFlow-Lite mobile application can also be adapted to other types of EEG devices for real-time applications.

1) Neural Networks: Temporal networks such as end-to-end Hierarchical Recurrent Neural Networks (RNN) and Bi-directional Long-Short Term Memory (LSTM) have been developed for classifying sleep stages from raw EEG in the works of SeqSleepNet [19] and DeepSleepNet [27] respectively. Though state of the art accuracy is achieved, they use multiple epochs of EEG (2.5 minutes and 12.5 minutes respectively) which impedes the time-resolution of labeling EEG. We use Time-distributed CNN to label 30 seconds of non-overlapping EEG. Our model is based on Time-Distributed

CNN [12] and is inspired by the DeepSleepNet from Supratak et al. [27]. DeepSleepNet makes use of representation learning with a CNN followed by sequence residual learning using Bidirectional-Long Short Term Memory cells (Bi-LSTM). The major drawback of this network is that it requires 25 epochs of raw EEG data to be fed in together to obtain 25 labels. This is mainly because of the Bi-LSTM which relies on large temporal sequences to achieve better accuracy.

SeqSleepNet [19] uses multiple epochs and outputs the sleep labels all at once using end-to-end Hierarchical RNN. It uses all 3 channels—namely, EEG, EMG and EOG in order to give the best overall accuracy of 87.1% on the MASS dataset [17]. CNN models by Sors et al. [25] and Tsinalis et al. [29], as well as SeqSleepNet and DeepSleepNet all use longer temporal sequences for inference—4, 5, 10 and 25 raw EEG epochs of 30 seconds respectively. We overcome this limitation by using Time-Distributed CNN to predict single 30-second real-time epochs using 1-channel EEG from Muse headband.

Further developing upon previous work [5], [16], [28] on spectral characteristics of EEG, we propose an improved on-device metric that enhances the identification of deep sleep (N3) using relative power spectral analysis from the EEG headband and a way of verifying the metrics that identify it.

Fig. 1. Model architecture used for training and inference

III. IMPLEMENTATION

A. Dataset description and pre-processing

We used the expanded Sleep-EDF database from Physionet-bank [9]. Single-channel EEG (Fpz-Cz at 100Hz) of 20 subjects is divided into a training-set of 33 nights and validation-set of 4 nights. Together, they contain non-overlapping nights of 19 subjects for 20-fold cross validation. The non-overlapping test-set contains 2 nights (1 subject). We remove the extra wake stages before and after half an hour of sleep as described in the DeepSleepNet [27]. We excluded *MOVE-MENT* and *UNKNOWN* stages, and combined N4 and N3 to follow 5-stage classification as per the new AASM rules [4].

B. Model architecture and training on SleepEDF dataset

Our model architecture is described in Figure 1. The Base-CNN has 3 repeated sets of two 1-D convolutional (Conv1D) layers, 1-D max-pooling and spatial dropout layers. This is followed by two Conv1D, 1-D global max-pooling, dropout and dense layers. We finally have a dropout layer as the output of Base-CNN. 30-second epochs of normalized EEG at 100Hz is fed into the Time-Distributed Base-CNN model [12] as described in Figure 1. All Conv1D layers use Rectified-Linear-Units (ReLU) activation. The training uses an Adam optimizer of 0.001 with an initial learning rate of e^{-3} which is reduced each time the validation accuracy plateaus using ReduceLROnPlateau Keras Callbacks.

C. Real-time EEG adaptation for testing on Muse

1) Pre-processing of EEG from Muse: Real-time brain activity from the flexible EEG headband is streamed via BLE to the phone. The raw EEG from Tp10 electrode is chosen for classification because the frontal electrodes (Af7 and Af8) are highly susceptible to noise from the EOG artefacts due to eye-movement. 30 seconds of raw EEG from Tp10 electrode is first notch-filtered with a central frequency of 60 Hz with a bandwidth of 10 Hz to remove power-line electrical disturbances and then with a band-pass filter of 1 Hz to 45 Hz because the starting frequency of delta sub-band is 1Hz and upper limit of gamma sub-band as described for the Muse EEG is 44Hz. We further down-sampled the filtered EEG to 100 Hz followed by another pass through the same band-pass filter to avoid noise from seeping in. The type of all filters used is IIR Butter-worth. 100 Hz of filtered EEG is then mean-shifted and divided by a constant scaling factor of 8. We then clip the normalized, filtered EEG spikes that exceeds the amplitude range [-4,4] in order to avoid K-complex like spikes caused by minor external motion artefacts.

Since the EEG device used for training is different from the testing one, we normalized the data by the wake-stage standard deviation of EEG (SD). We estimated the mutual information for a discrete sleep stage variable with a set of statistical EEG features—Median, SD, Range, Variance, Mean, Skew and MMD [1] that is calculated every 30 seconds as input and the corresponding sleep-stage labels as output. Using the Mutual-Classif-Info function of Scikit-Learn Machine Learning library, we calculated this statistical feature importance of EEG for 3 randomly selected nights. SD clearly plays major role with average relative importance of 53.33 percent over other features for the classification of sleep stages. Hence, normalization using wake stage SD for 30-second epoch adaptation helps in making EEG both instrument-independent and subject-independent as long as signal is noise-reduced.

D. Spectral Conditioning of EEG

1) Power spectral bands from the Muse EEG: Power spectral analysis of EEG gives the 5 major sub-bands of brain-waves, namely, alpha, beta, gamma, delta and theta. The corresponding frequency bins are specified by Muse [13]. If absolute band powers of given frequency sub-bands are calculated as the sum of power spectral density (PSD) of EEG over that frequency range, then relative band powers are calculated by dividing absolute band power of that particular sub-band over sum of absolute powers in all sub-bands. Frequency resolution of the bins for calculating PSD is 0.86 Hz per bin with 128 bins giving 110 Hz as the maximum frequency that Muse (sampling rate: 220Hz) can detect.

2) Power spectral measures for identifying Deep Sleep: Deep Sleep is characterized by its dominant high-amplitude low frequency delta waves. Out of the 5 sub-bands of EEG, delta power is the highest in amplitude while gamma amplitude is the least during deep sleep. The ratio of band powers (Deep Ratio = δ/γ), the sum of delta power divided by sum of gamma power over 30 seconds and multiplied by a scaling

factor (0.1), is used as the measure for the depth of sleep. Different ratios such as δ/α, δ/β, δ/θ and $(\alpha + \beta)/(\delta + \theta)$ have been explored in previous works [28] [6]. Our measure δ/γ represents a sharper contrast between deep sleep and other stages of sleep as compared to the previously mentioned measures. The significant drop in the amplitude of this ratio also clearly marks the end transition of deep sleep.

3) Verification module for the power spectral measures: In order to visualize and verify the power spectral bands and their ratios, we developed a module that calculates the power bands of EEG form SleepEDF according to the Muse specifications [10]. Welch's method from the open-source MNE Python package is used to calculate the PSD of 100 Hz EEG with the specified 0.86 Hz/bin frequency resolution and the corresponding frequency ranges of EEG sub-bands. We then calculated the ratio δ/γ over non-overlapping 30 seconds of EEG. We repeated this method to visualize 4 randomly picked nights from the Sleep EDF dataset. Further, we prepared a classification report of deep sleep given a particular threshold for δ/γ to test the efficacy of our measure. In the direction of designing a calibration method for calculating the corresponding subject-dependent deep-sleep threshold while the subject is awake, we also consider the relationship between the δ/γ thresholds that give the best accuracy for classification of deep sleep and the corresponding mean of δ/γ values during wake-stage for 4 test nights of sleep.
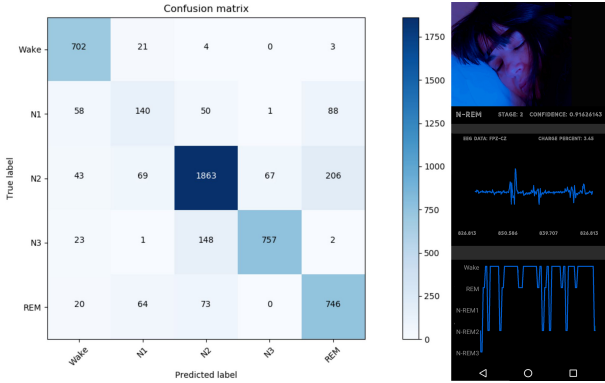

Fig. 2. Left: Confusion matrix. Right: Person sleeping with the flexible EEG (AF7, AF8, TP9, TP10 and reference Fpz) along with our android app

IV. RESULTS

A. Evaluation on the SleepEDF Dataset

Our model has an overall accuracy of 83.5% for 20-fold cross validation of 5-stage classification. The accuracy for nights from the test data ranges from 72%(worst-case). This model achieves reliable accuracy given that the overall IRR (Inter-Rater-Reliability) [8] among human experts scoring sleep recordings reported was about 80% (Cohens κ = 0.68 to 0.76). Table I describes the precision, recall, F1-score and support of all the 5 sleep stages on predictions from 5 test nights. The corresponding accuracy of 81.72% and F1-score of 76.23% was obtained. The confusion matrix for the same is

TABLE I
5-CLASS CLASSIFICATION REPORT FOR SLEEP STAGING ON 5 TEST NIGHTS

Label	Sleep Stage	Precision	Recall	F1-score	Support
0	Wake	0.83	0.96	0.89	730
1	N1	0.47	0.42	0.44	337
2	N2	0.87	0.83	0.85	2248
3	N3	0.92	0.81	0.86	931
4	REM	0.71	0.83	0.77	903
Micro	average	0.82	0.82	0.82	5149
Macro	average	0.76	0.77	0.76	5149
Weighted	average	0.82	0.82	0.82	5149

shown in the left part of Figure 2. N1 stage shows the poorest agreement because of the absence of occipital electrode [11]. Figure 3 shows hypnograms of a full-night predicted by the model and the ground truth as labeled in the dataset.
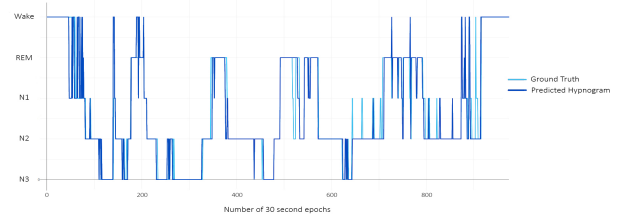

Fig. 3. Comparison of a full night ground truth and predicted hypnograms

B. Deep Learning Model Predictions on Muse EEG

Figure 4 shows the predicted hypnogram of a test subject on our smartphone app using just a Tp10 Muse channel. It consists of 2 cycles of sleep with 1st and 2nd deep sleep onsets around 25 minutes and 2 hours from the start respectively.
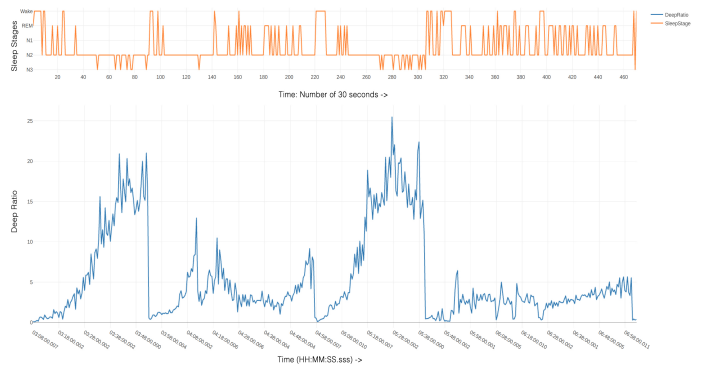

Fig. 4. Hypnogram predicted by our app along with δ/γ visualization

C. External Power Spectral Conditioning for Deep Sleep

The ratio: δ/γ over non-overlapping 30 seconds of EEG is also plotted against its hypnogram in Figure 4. This ratio peaks with high amplitude during deep sleep. The end-transition is also characterized distinctively by the sudden large amplitude drop. This provides an alternative way of detecting deep sleep by power spectral analysis. Deep sleep classification accuracy varies with the threshold value of δ/γ. The mean of δ/γ during

wake stage [0.023, 0.113, 0.174, 0.400] is in non-decreasing relationship with the final thresholds [0.7, 0.7, 3.2, 4.2] that gives maximum accuracy [96.5%, 97.9%, 97.2% and 97.3%] on 4 test nights of sleep respectively from the dataset.

V. Conclusions and Future Work

This work demonstrates an end-to-end mobile pipeline for the first real-time sleep-staging of a wearable EEG with the fastest real-time resolution of 30 seconds. The non-decreasing relationship of δ/γ during wake stage and deep sleep can be further used to design subject-dependent calibration methods so as to set the thresholds for automatic identification of deep sleep. Moreover, our automatic sleep staging algorithm does not require external servers and can be used outside of hospitals and research laboratories. The app is versatile as it can be adapted to take in single channel recordings from any wearable EEGs. We aim to use this work for real-time interventions using Brain Computer Interfaces (BCI) for applications in HCI. Specifically, we hope to integrate wearable olfactory systems [2] that can automatically intervene in real-time during specific sleep stages to enhance the quality of sleep along with real-time audio-neural feedback [24] and sleep-based enhancement of learning and memory [3], [21].

References

[1] Khald Ali I Aboalayon, Miad Faezipour, Wafaa S Almuhammadi, and Saeid Moslehpour. Sleep stage classification using eeg signal analysis: a comprehensive survey and new investigation. *Entropy*, 18(9):272, 2016.

[2] Judith Amores, Javier Hernandez, Artem Dementyev, Xiqing Wang, and Pattie Maes. Bioessence: A wearable olfactory display that monitors cardio-respiratory information to support mental wellbeing. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5131–5134. IEEE, 2018.

[3] Thomas Andrillon, Daniel Pressnitzer, Damien Léger, and Sid Kouider. Formation and suppression of acoustic memories during human sleep. *Nature communications*, 8(1):179, 2017.

[4] Richard B Berry, Rita Brooks, Charlene E Gamaldo, Susan M Harding, CL Marcus, BV Vaughn, et al. The aasm manual for the scoring of sleep and associated events. *Rules, Terminology and Technical Specifications, Darien, Illinois, American Academy of Sleep Medicine*, 2012.

[5] Christian Berthomier, Xavier Drouot, Maria Herman-Stoïca, Pierre Berthomier, Jacques Prado, Djibril Bokar-Thire, Odile Benoit, Jérémie Mattout, and Marie-Pia d'Ortho. Automatic analysis of single-channel sleep eeg: validation in healthy individuals. *Sleep*, 30(11):1587–1595, 2007.

[6] Eugene N Bruce, Margaret C Bruce, and Swetha Vennelaganti. Sample entropy tracks changes in eeg power spectrum with sleep state and aging. *Journal of clinical neurophysiology: official publication of the American Electroencephalographic Society*, 26(4):257, 2009.

[7] Alexander J Casson, David C Yates, Shelagh JM Smith, John S Duncan, and Esther Rodriguez-Villegas. Wearable electroencephalography. *IEEE engineering in medicine and biology magazine*, 29(3):44–56, 2010.

[8] Heidi Danker-hopfe, Peter Anderer, Josef Zeitlhofer, Marion Boeck, Hans Dorn, Georg Gruber, Esther Heller, Erna Loretz, Doris Moser, Silvia Parapatics, et al. Interrater reliability for sleep scoring according to the rechtschaffen & kales and the new aasm standard. *Journal of sleep research*, 18(1):74–84, 2009.

[9] Bob Kemp, Aeilko H Zwinderman, Bert Tuk, Hilbert AC Kamphuisen, and Josefien JL Oberye. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the eeg. *IEEE Transactions on Biomedical Engineering*, 47(9):1185–1194, 2000.

[10] Olave E Krigolson, Chad C Williams, Angela Norton, Cameron D Hassall, and Francisco L Colino. Choosing muse: Validation of a low-cost, portable eeg system for erp research. *Frontiers in neuroscience*, 11:109, 2017.

[11] Brendan P Lucey, Jennifer S Mceland, Cristina D Toedebusch, Jill Boyd, John C Morris, Eric C Landsness, Kelvin Yamada, and David M Holtzman. Comparison of a single-channel eeg sleep study to polysomnography. *Journal of sleep research*, 25(6):625–635, 2016.

[12] Youness Mansar. EEG_classification, 2018.

[13] Muse. *Available Data: Muse specifications*, 2016.

[14] Takashi Nakamura, Valentin Goverdovsky, Mary J Morrell, and Danilo P Mandic. Automatic sleep monitoring using ear-eeg. *IEEE journal of translational engineering in health and medicine*, 5:1–8, 2017.

[15] Anh Nguyen, Raghda Alqurashi, Zohreh Raghebi, Farnoush Banaei-Kashani, Ann C Halbower, Thang Dinh, and Tam Vu. In-ear biosignal recording system: A wearable for automatic whole-night sleep staging. In *Proceedings of the 2016 Workshop on Wearable Systems and Applications*, pages 19–24. ACM, 2016.

[16] Julie A Onton, Dae Y Kang, and Todd P Coleman. Visualization of whole-night sleep eeg from 2-channel mobile recording device reveals distinct deep sleep stages with differential electrodermal activity. *Frontiers in human neuroscience*, 10:605, 2016.

[17] Christian O'reilly, Nadia Gosselin, Julie Carrier, and Tore Nielsen. Montreal archive of sleep studies: an open-access resource for instrument benchmarking and exploratory research. *Journal of sleep research*, 23(6):628–635, 2014.

[18] Amiya Patanaik, Ju Lynn Ong, Joshua J Gooley, Sonia Ancoli-Israel, and Michael WL Chee. An end-to-end framework for real-time automatic sleep stage classification. *Sleep*, 41(5):zsy041, 2018.

[19] Huy Phan, Fernando Andreotti, Navin Cooray, Oliver Y Chén, and Maarten De Vos. Seqsleepnet: End-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging. *arXiv preprint arXiv:1809.10932*, 2018.

[20] Djordje Popovic, Michael Khoo, and Philip Westbrook. Automatic scoring of sleep stages and cortical arousals using two electrodes on the forehead: validation in healthy adults. *Journal of sleep research*, 23(2):211–221, 2014.

[21] Björn Rasch, Christian Büchel, Steffen Gais, and Jan Born. Odor cues during slow-wave sleep prompt declarative memory consolidation. *Science*, 315(5817):1426–1429, 2007.

[22] A Sano, AJ Phillips, AW McHill, S Taylor, LK Barger, CA Czeisler, and RW Picard. 0182 influence of weekly sleep regularity on self-reported wellbeing. *Journal of Sleep and Sleep Disorders Research*, 40(suppl_1):A67–A68, 2017.

[23] Nicolas Schaltenbrand, Régis Lengelle, M Toussaint, R Luthringer, G Carelli, A Jacqmin, E Lainey, Alain Muzet, and Jean-Paul Macher. Sleep stage scoring using the neural network model: comparison between visual and automatic analysis in normal subjects and patients. *Sleep*, 19(1):26–35, 1996.

[24] Maren D Schütze and Klaus Junghanns. The difficulty of staying awake during alpha/theta neurofeedback training. *Applied psychophysiology and biofeedback*, 40(2):85–94, 2015.

[25] Arnaud Sors, Stéphane Bonnet, Sébastien Mirek, Laurent Vercueil, and Jean-François Payen. A convolutional neural network for sleep stage scoring from raw single-channel eeg. *Biomedical Signal Processing and Control*, 42:107–114, 2018.

[26] Andreas Stochholm, Kaare Mikkelsen, and Preben Kidmose. Automatic sleep stage classification using ear-eeg. In *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*, pages 4751–4754. IEEE, 2016.

[27] Akara Supratak, Hao Dong, Chao Wu, and Yike Guo. Deepsleepnet: a model for automatic sleep stage scoring based on raw single-channel eeg. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(11):1998–2008, 2017.

[28] K Šušmáková and A Krakovská. Classification of waking, sleep onset and deep sleep by single measures. *Sci. Rev*, 7:34–38, 2007.

[29] Orestis Tsinalis, Paul M Matthews, and Yike Guo. Automatic sleep stage scoring using time-frequency analysis and stacked sparse autoencoders. *Annals of biomedical engineering*, 44(5):1587–1597, 2016.

[30] Jin Zhang, Dawei Chen, Jianhui Zhao, Mincong He, Yuanpeng Wang, and Qian Zhang. Rass: A portable real-time automatic sleep scoring system. In *Real-Time Systems Symposium (RTSS), 2012 IEEE 33rd*, pages 105–114. IEEE, 2012.

[31] Mingmin Zhao, Shichao Yue, Dina Katabi, Tommi S Jaakkola, and Matt T Bianchi. Learning sleep stages from radio signals: a conditional adversarial architecture. In *International Conference on Machine Learning*, pages 4100–4109, 2017.