A Deep Learning Approach for Detecting Obstructive Sleep Apnea from Wearable Technology

Oluseye Bankole
Cornell Tech
ob97@cornell.edu

Jake Bass
Cornell Tech
jab783@cornell.edu

Matthew Dalton
Cornell Tech
mgd67@cornell.edu

Travis Allen
Cornell Tech
twa24@cornell.edu

1 Introduction

Obstructive Sleep Apnea (OSA) is a sleep disorder that is characterized by intermittent blockage of upper respiratory airways during sleep. If left untreated, OSA can lead to serious chronic conditions such as cardiovascular disease and metabolic dysfunction. While there are a number of treatments available, an estimated 80% of cases go undiagnosed and therefore untreated. [1]

The current gold standard for diagnosis is expert annotation of Polysomnography (PSG), otherwise known as a sleep study. In a PSG, multiple signals such as EEG (electroencephalogram), ECG (electrocardiogram), EMG (electromyogram), airflow, oxygenation in arterial blood, etc. are collected over a night's sleep. Using PSGs, experts can diagnose OSA with high accuracy. However, PSGs are expensive, time-consuming, and labor-intensive, which can largely account for why so many individuals go undiagnosed [2]. There clearly exists a need to simplify the diagnostic process.

Over the last five years smart-watch adoption has grown considerably, and with it a new way for people to track their health patterns, such as heart rate and skin temperature. The abundance of this passively-collected data has driven research into using smartwatch activity to make various diagnoses [3]. However, an area that remains unexplored is using smartwatch signals to diagnose OSA.

In this work, we attempt to lay the foundation for an apnea diagnostic that would rely solely on data from wearables. While there is some successful scholarship on automated diagnostic methods, expert annotation is still the gold standard, and with good reason. There are many challenges in creating an effective diagnostic, and even more in translating it to a wearable device. Not only are there far fewer types of sensor data available (an Apple Watch certainly does not take brain wave readings), the data that does exist is of much lower fidelity. Instead of the hundreds of readings taken per second in a PSG, the Apple Watch samples at a rate of .7 Hz. This work focuses on exploring the viability of diagnostic systems that rely on data at the lower granularity seen in wearables. We reproduce a baseline LSTM-based model that works with features derived solely from ECG data, and then evaluate variations of this model on data that more closely resembles wearable data. (Note that wearables contain beats per minute heart rate data, which can be viewed as a lower fidelity ECG signal.) In reproducing the baseline, we achieved an AUC on the test set of 0.7856. In our evaluations of the model in the proxy wearable setting, we achieved AUC scores on test of .604 and .6672. We hope our work paves the way for diagnostic methods that can run on wearables, which we hope increase accessibility to testing and decrease the population that unknowingly suffers from obstructive sleep apnea.

This project is part of a larger research collaboration between Northwell Health and Cornell Tech. In addition to the students named in this paper, Michael Kirschenbaum, a Psychiatry Resident at Northwell, and Prathamesh Param Kulkarni, a postdoc fellow at Northwell and Cornell Tech, are working on it. Additionally, it is part of independent research project for Jake Bass advised by Professor Deborah Estrin.

2 Related Work

Some of the earlier works in this area focused on reducing the number of channel signals from the PSG necessary to diagnose apnea. [4] evaluates the fidelity of using a single channel, the ECG signal, in identifying moderate to high sleep apnea (SA) in sleep subjects. The authors use two algorithms, namely cardiopulmonary coupling (CPC) and cyclic variation of heart rate (CVHR). The results from the two algorithms were both compared against manual scoring, the current method of evaluating sleep apnea, and were found to show strong agreement with the gold standard. Their study showed that it is possible to get reasonable results from using a single ECG channel.

In the last few years, there has been interest in using deep learning techniques to detect sleep apnea from single ECG signal inputs. [5] pass the R-R interval – the difference between successive R peaks from the ECG signal – into a LSTM network, culminating in a softmax classifier. The authors show that the this LSTM-based model predicts obstructive sleep apnea (OSA) better than some of the traditional simple neural network approaches that fell far short of the gold standard.

Rather than using feature engineering approaches that relied on knowledge of ECG signals or of OSA, [6] used auto encoders to learn the most relevant representation of the data. To do this, they took the unlabeled RR interval of the ECG signal and fed into a deep neural network, from which they learnt the auto-encoded representation. They also evaluated the temporal dependence of the signal segments using a Hidden Markov Model (HMM). Using this methodology, the authors were able to achieve a sensitivity of 88.9%, which is higher than values previously reported in the literature.

Given the limitation on the accuracy and specificity of some of the earlier deep learning methods, other authors have tried to explore newer deep learning techniques such as convolutional networks. For example, [2] attempted a convolutional neural network to detect OSA from single lead ECG signals. The idea is that the convolution layers supposedly learn features such as the temporal variation across the input signal and the regions of high amplitude in the data. In the pooling layer, rather than use simple pooling techniques like max pooling, average pooling or l2-norm pooling, the authors add another convolutional layer but apply a separate slide for the time dimension during kernel shifting. The authors were able to achieve accuracy and sensitivity values of 98.9% and 97.8% respectively, which significantly trounces results from previous studies.

Given the scarcity of equipment that is able to capture ECG data, there has been interest among other researchers in investigating whether heart rates can be used as a proxy for ECG input signal. [7] use instantaneous heart rate (IHR) with a deep network to detect severe sleep apnea. It is important to note that the IHR data used in this paper was derived from ECG data measurements using a toolkit. For model training, they employ an LSTM-based architecture to learn the dynamics of the IHR dataset. Their results showed accuracies greater than 85% accuracy, which is comparable to results from other studies using ECG input signals.

3 Datasets

We utilize data from two sources: Physionet, a large open-source collection of recorded physiologic signals, and the National Sleep Research Resource (NSRR), which offers "free access to large collections of de-identified physiological signals and clinical data elements collected in well-characterized research cohorts and clinical trials." [9; 11] We chose these sources because they were open-source, designed for researchers, and had datasets that had apnea annotations and a full bank of PSG sensor data that would allow us to replicate wearable data. We intended to use proprietary data from our collaboration with Northwell, but due to delays in the study, there is not a significant amount ready for use at this time.

3.1 Physionet

Physionet provides one dataset that we used: the CINC 2000 dataset.[10] This dataset was developed for a cardiology competition where researchers attempted to screen and quantify obstructive sleep apnea using ECG-based methods. The screening task called for grouping patients into three classes:

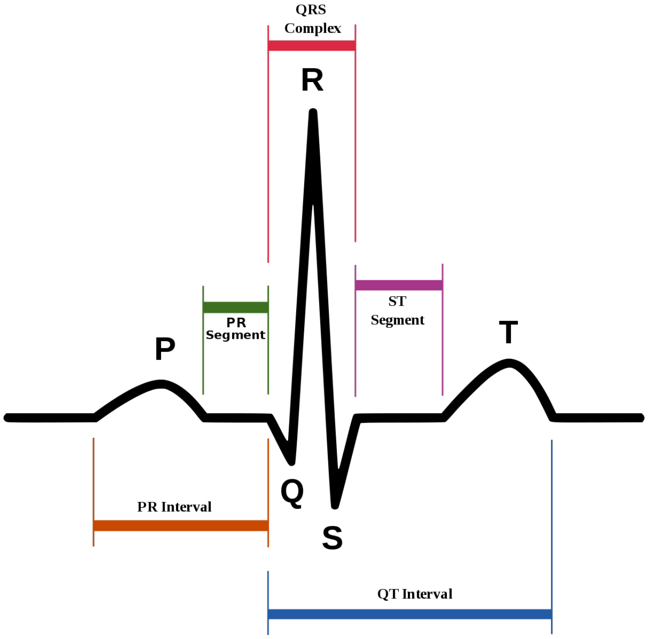


Figure 1: An annotated QRS complex.

apneic, borderline, and control. The detection task required classifying each minute as apneic or normal. Expert annotations were provided for each minute as ground truth.

The datasets comes with train and test sets, each with 35 patients recorded for a whole night. In both sets there are 20 apneic patients (class A), 5 borderline patients (class B), and 10 control patients (class C). These splits are defined in part based on Apnea-Hypopnea Index (AHI). AHI is a measure of apneic events per hour over the course of the night and is traditionally used to classify severity of apnea: for adults fewer than 5 events is non-apneic, 5 - 15 indicates mild apnea, 15 - 30 indicates moderate apnea, and more than 30 indicates severe apnea. Class A recordings must contain at least one hour of sleep with an AHI of at least 10 and must have at least 100 minutes classified as apneic over the course of the night. Class B recordings have at least one hour of sleep with an AHI over 5 and between 5 and 99 apneic minutes. Class C recordings must have fewer than 5 minutes of apnea over the course of the night.

The competition intended for researchers to utilize ECG-based feature representations. In addition to the raw ECG signal, they released QRS annotations. A QRS annotation, seen above, is the name of the complex seen in the electrical signal of a heartbeat, and is commonly used to calculate beats per minute.

3.2 National Sleep Research Resource

The NSRR provides one dataset that we used: the Apnea, Bariatric surgery, and CPAP study (ABC) Dataset.[12] The ABC dataset was taken from a study aimed at assessing the effects of two different treatments 1) bariatric (weight loss) surgery and 2) continuous positive airway pressure (CPAP) therapy plus weight loss counseling on patients with with class II obesity and those who have severe obstructive sleep apnea (OSA).

This dataset contains 49 adults with severe obstructive sleep apnea. Each patient was recorded before treatment, and then at 9 and 18 months after treatment as a follow-up. It contains full polysomnography data from each recording, along with expert apnea annotations. Importantly, the PSG includes ECG, along with pulse data measured directly from a pulse oximeter, as opposed to derived via the QRS complex.

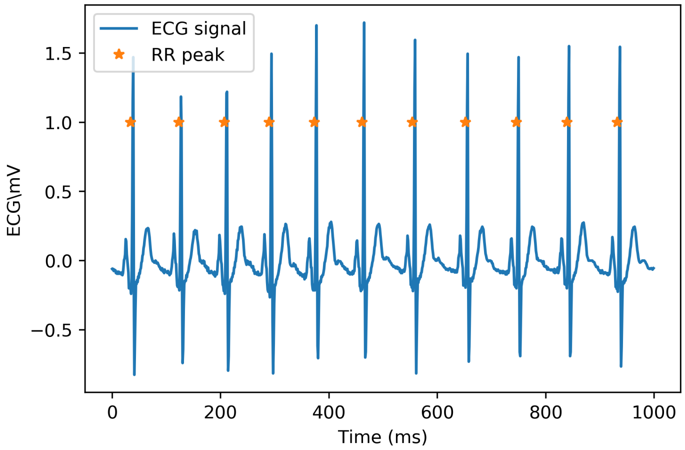


Figure 2: Raw ECG data for Record A01 (Physionet) with RR peak overlay

4 Methodology

4.1 Recreating Baseline Model

To reproduce the results reported by [7], we collected the physionet dataset and pre-processed it to derive the features used by Vinyakumar et. al. This dataset contains ECG recordings sampled at 100 Hz, sleep annotations every 1 minute and QRS detection annotation times. Figure 2 shows the ECG signals with the QRS annotations for a sample record, with the time annotations occurring right where the ECG peak signal occurs.

4.1.1 Preprocessing Data (ECG to IHR)

We first converted the ECG signals to IHR using a QRS converter from the WFDB package but very quickly noticed that the IHR signals were unstable. As a result, we used the QRS annotations from the physionet dataset to extract out the RR intervals. The RR intervals (in minutes) were converted to IHR signals (beats per minute) using the following formula:

$$IHR = \frac{60}{RR_{interval}}$$

More importantly, the RR intervals were determined by taking the difference between successive QRS peaks.

Only the first 30 beats in each minute recording are considered during our analysis. Minutes with fewer than 30 beats are discarded. This treatment is consistent with the method used by [7].

4.1.2 Model Architecture

In an effort to reproduce the results published by [7], we built a stacked LSTM-RNN in accordance with the architecture described in their paper. The network consists of three layers: an input layer, a hidden layer, and an output layer. The input layer is comprised of 30 neurons for the 30 IHR-per-minute recordings in our physionet dataset. The hidden layer consists of 32 memory units in the LSTM cell, although we also experiment with this hyperparameter. Finally, we add an output layer with a sigmoid activation function for our binary classification of "Apnea" or "Non-Apnea". In addition to the layers described in the paper, we also implemented dropout regularization by adding a dropout layer which randomly dropped 10% of neurons during training.

Binary cross-entropy loss was used as our loss function, and Adam optimization was used to iteratively update our network weights during training. Hyperparameters that we experimented with during training included the number of epochs as well as the size of the training split.

The model was implemented in Keras.

4.2 Testing Baseline Model on Pulse Data

Our next step was to test the baseline model on pulse data from the NSRR ABC dataset, in order to validate the trained model. The purpose of this validation was twofold. First, we wanted to ensure that the model generalizes to beyond the relatively small Physionet dataset. Second, we wanted to assess how well the baseline model performs on directly observed pulse data as opposed to the ECG-derived IHR signals it was trained on. This is important because it serves as a proxy for how the model would perform "in the wild", for example, deployed on a wearable with access only to pulse data and not an ECG feed.

The pulse data from the ABC dataset is sampled at a rate of 16 Hz; so, for a single minute, we have 960 observations of the subject's pulse rate. These readings are too granular compared with what is expected from a wearable, so we downsample the pulse rate observations to 1 Hz, or 60 observations per minute. Downsampling is done via fast Fourier transform (FFT) as implemented in SciPy. Furthermore, the baseline model takes in fixed-length input vectors of length 30, so for each minute, we use the first 30 downsampled observations as our feature vector. Performing these operations gives us test examples that represent a minute of sleep with apnea annotations and a 30 dimensional feature vector. For evaluation, we test the trained model on a 2188 example test set derived from 5 ABC records. We also report results on a smaller 1562 example test set with balanced apneic and non-apneic classes, since some of the metrics we report can be skewed by class imbalance. For each test set, we report accuracy, precision, recall, F1-score, and AUC, which can be found in Section 5.2.

4.3 Creating a Pulse Baseline Model

In addition to the previous experiments, we created a model that was trained on pulse data. While it would be possible to train on the extracted IHR data and deploy that model to a wearable, ideally we would like to train on data that is most similar to data from a wearable's sensors. This would likely improve performance through increased train-test similarity, and would allow us to more easily incorporate data gathered from wearables into new models. To enact this experimental condition, we took the feature representation described in Section 4.2 and trained on it. Since we truncate the data to a fixed length of 30, we are able to use the same architecture as we did previously. We used 8828 annotated minutes from the NSRR dataset drawn from 20 subjects with a 70-10-20 train-val-test-split. We trained the model for 1000 epochs. Results are reported in Section 5.3 for both the full and balanced test sets.

5 Results

5.1 Recreating Baseline Model

Hyperparameters	Accuracy	Recall	Precision	F1 Score	AUC
1000 Epochs, 4 Memory Units 50-10-40 train-val-test split	0.7264	0.4775	0.7103	0.5711	0.7593
1000 Epochs, 32 Memory Units 50-10-40 train-val-test split	0.7370	0.5750	0.6851	0.6252	0.7743
5000 epochs, 32 Memory Units, 80-10-10 train-val-test split	0.75	0.5997	0.7016	0.6466	0.7856

Vinyakumar et. al claimed results that yielded an AUC of 0.99, with a corresponding accuracy, precision, and recall of 89.0%, 82.4%, and 99.4% respectively [7]. Despite following the implementation described, recurring correspondence with the authors, and extensive hyperparameter experimentation, we were not able to reproduce the authors' results. We were able to locate an open-sourced repository from the Vinyakumar et. al with the model definition. However, we did not have their code for preprocessing or evaluation. As such, we identify the data representation and the dataset imbalance corrections as possible areas where our implementation differs from theirs. In particular, we believe there are better ways to represent the data than the truncate thirty beats methodology, and that noise inherent to the QRS annotation process could account for differences in our results. Additionally, Vinyakumar et. al say "class distribution details of A and N are omitted due to paucity of space". Since the classes are quite imbalanced, any corresponding modifications to training set balance or evaluation procedure to account for the imbalance could dramatically affect the results. We tried various sampling procedures to account for this imbalance, but Vinyakumar et. al's procedures could differ. Consequently, we view our baseline as a more reasonable point of comparison.

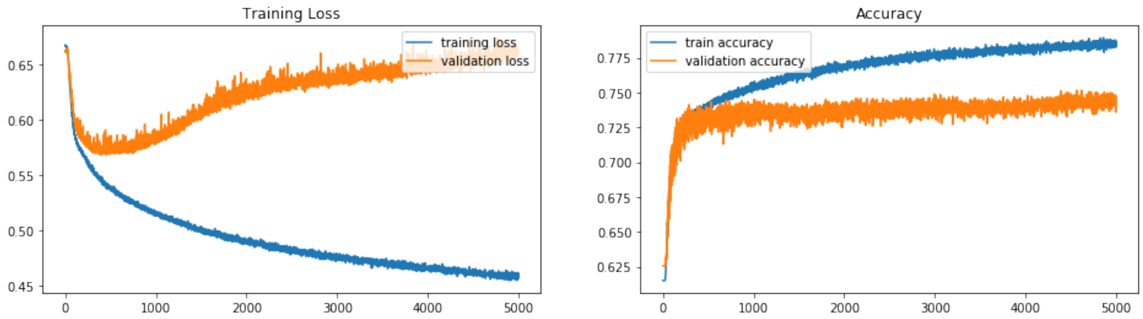


Figure 3: Recreating Baseline: Training Loss And Accuracy

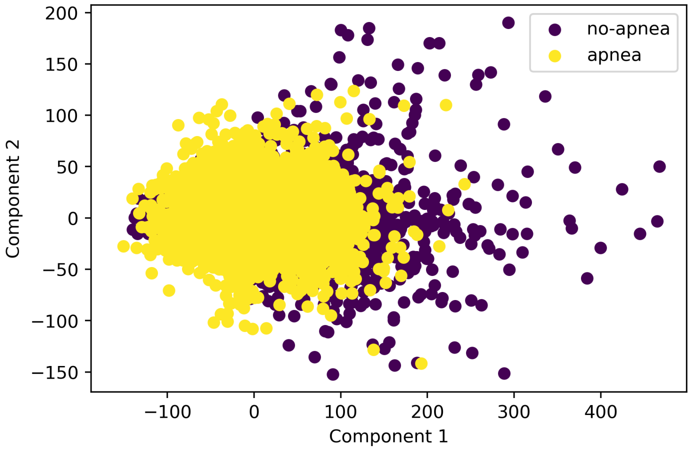
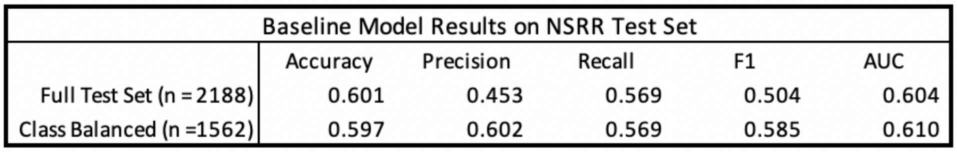


Figure 4: PCA decomposition of physionet data with 30 IHR beat length

Despite the fact that we did not reproduce the authors' results, our network *does* learn something meaningful. Given that that the True Negative:True Positive ratio is 60:40 in our test set, our highest-performing model (the last row in the preceding table) is significantly outperforming chance and has a respectable .7856 AUC metric. In addition, figure 4 shows that after performing PCA to reduce our data to 2 dimensions, the apneic and non-apneic minutes are not easily separable. Although we did perform dimensionality reduction, this nonetheless highlights the inherent difficulty in discriminating these two classes and therefore supports our results. Thus, this model serves as a valuable baseline for our further experiments whose results are discussed below.

5.2 Testing Baseline Model on Pulse Data

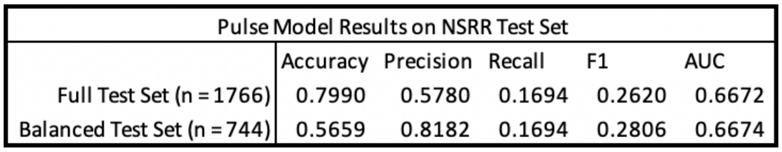


Baseline Model Results on NSRR Test Set

	Accuracy	Precision	Recall	F1	AUC
Full Test Set (n = 2188)	0.601	0.453	0.569	0.504	0.604
Class Balanced (n =1562)	0.597	0.602	0.569	0.585	0.610

Evaluating the baseline model on the NSRR data gives us some sense of how this model would perform if it were to be deployed and used on direct pulse data. We see that on this new data, the model does not perform as well, achieving a .604 AUC as compared with .7856 AUC on the Physionet dataset. This performance drop is to be expected as we are switching from heart rate data derived from highly informative ECG signals to the cruder pulse signals at test time. Nevertheless, an AUC metric of .604 does indicate that the baseline model is beating random chance and does have some ability to generalize to real-world pulse-data.

5.3 Creating a Pulse Baseline Model



Pulse Model Results on NSRR Test Set

	Accuracy	Precision	Recall	F1	AUC
Full Test Set (n = 1766)	0.7990	0.5780	0.1694	0.2620	0.6672
Balanced Test Set (n = 744)	0.5659	0.8182	0.1694	0.2806	0.6674

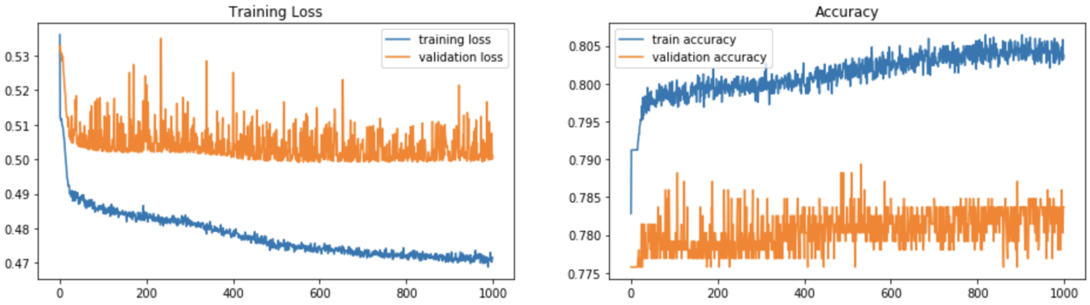


Figure 5: Creating a Pulse Baseline Model: Training Loss And Accuracy

Our pulse baseline model takes the architecture described in Vinyakumar et. al [7], and trains it on direct pulse rate data from NSRR instead of the ECG-derived IHR data that was originally used. We then test this model on data also drawn from NSRR. We report an AUC of .6672 on the test set (full results listed above). Again, while these results are significantly below what was originally reported in Vinyakumar et. al [7], they do provide evidence toward the potential of this model to generalize. Additionally, these results are better than those seen in Section 5.2. While this result is expected given that this model has more similar train and test representations than our prior one, this result is still significant. We view this performance as evidence that a model trained on wearable-level sampling frequencies can successfully identify apnea.

6 Conclusion

Overall, our report contains three significant conclusions. The first was our inability to reproduce the remarkable results obtained in prior work, as discussed in Section 5.1. We discuss how this result may be because of inadequacies in their reporting of preprocessing procedures and class imbalance adjustments. Despite this lack of information, we were able to create a viable baseline that performs substantially better than random chance.

Secondly, we found in Section 5.2 that we can obtain above-average results for real pulse data at wearable-level sampling frequency evaluated on a model trained on ECG-derived IHR data. Since the performance of the pulse test set is slightly worse than the IHR test set, we expect that augmenting the training dataset with pulse data might improve model performance. Additionally, this exact performance of this experiment is not significant; its purpose was to function as a proof of concept for our final experiment.

Lastly, we found in Section 5.3 that a model trained on wearable-level sampling frequencies is able to successfully distinguish between apneic and non-apneic minutes. While the results are not as strong as previous work on higher fidelity data, our result carries increased clinical significance. Because apnea so commonly goes undiagnosed (more than 80 percent of cases in the US are undiagnosed), there is great need for a pre-screening procedure that would help funnel people into the more comprehensive and expensive gold standard PSG diagnostic. We hope that our results lay the groundwork for future research into a diagnostic that could run entirely on a wearable, thus greatly increasing diagnostic accessibility.

References

[1] Chelba, et al. “One Billion Word Benchmark for Measuring Progress in Statistical Language Modeling.” ArXiv.org, 4 Mar. 2014, arxiv.org/abs/1312.3005.

[2] Dey, Debangshu, et al. “Obstructive Sleep Apnoea Detection Using Convolutional Neural Network Based Deep Learning Framework.” Biomedical Engineering Letters, vol. 8, no. 1, 2017, pp. 95–100., doi:10.1007/s13534-017-0055-y.

[3] Li, Xiao, et al. “Digital Health: Tracking Physiomes and Activity Using Wearable Biosensors Reveals Useful Health-Related Information.” PLOS Biology, vol. 15, no. 1, 2017, doi:10.1371/journal.pbio.2001402.

[4] Magnusdottir, S., & Hilmisson, H. (2018). Ambulatory screening tool for sleep apnea: analyzing a single-lead electrocardiogram signal (ECG). Sleep and Breathing, 22(2), 421-429.

[5] Cheng, M., Sori, W. J., Jiang, F., Khan, A., & Liu, S. (2017, July). Recurrent neural network based classification of ECG signal features for obstruction of sleep apnea detection. In 2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC) (Vol. 2, pp. 199-202). IEEE.

[6] Li, K., Pan, W., Li, Y., Jiang, Q., & Liu, G. (2018). A method to detect sleep apnea based on deep neural network and Hidden Markov model using single-lead ECG signal. Neurocomputing, 294, 94-101.

[7] Pathinarupothi, R. K., Vinaykumar, R., Rangan, E., Gopalakrishnan, E., & Soman, K. P. (2017, February). Instantaneous heart rate as a robust feature for sleep apnea severity detection using deep learning. In 2017 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI) (pp. 293-296). IEEE.

[8] Pan, Jiapu, and Willis J. Tompkins. “A Real-Time QRS Detection Algorithm.” IEEE Transactions on Biomedical Engineering, BME-32, no. 3, 1985, pp. 230–236., doi:10.1109/tbme.1985.325532.

[9] Physionet. https://physionet.org/.

[10] Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation 101(23):e215-e220 [Circulation Electronic Pages; http://circ.ahajournals.org/content/101/23/e215.full]; 2000 (June 13). From https://physionet.org/physiobank/database/apnea-ecg/ via https://physionet.org/challenge/2000/#data-for-development-and-evaluation.

[11] The National Sleep Research Resource. https://sleepdata.org.

[12] Dean DA 2nd, Goldberger AL, Mueller R, Kim M, Rueschman M, Mobley D, Sahoo SS, Jayapandian CP, Cui L, Morrical MG, Surovec S, Zhang GQ, Redline S. Scaling Up Scientific Discovery in Sleep Medicine: The National Sleep Research Resource. Sleep. 2016 May 1;39(5):1151-64. doi: 10.5665/sleep.5774. Review. PubMed PMID: 27070134; PubMed Central PMCID: PMC4835314. From https://sleepdata.org/datasets/abc.