Identifying NREM Sleep Stages in Consumer Wearables
Final Paper

Presentation Link: https://www.youtube.com/watch?v=xS4Df2oXbMg
Github Link: https://github.gatech.edu/abutchers3/CSE-6250-Final-Project

Daniel J. Fasciano, BS[1], Matthew S. Shimko, BS[1], Tran N. Ton, BS[1], An T. Butchers, BS[1]
[1]Georgia Institute of Technology, Atlanta, GA, USA

Abstract

In this project, we worked with Apple Watch health data such as acceleration and heart rate gathered from June 2017 to March 2019 by the University of Michigan. The data came from 31 subjects. We built a sleep stage classifier based on an existing work by Walch et al[4, 9] that is available on PhysioNet[10]. While the original work has a Wake vs Sleep and Wake/NREM/REM sleep classifier, we successfully developed more advanced classifications with more classes classified such as Wake/N1/N2/N3/REM and Wake/N1+N2/N3/REM to bring more benefits to sleep studies[8]. We also tried AdaBoost to test performances of different algorithms.

Introduction

Sleep is an important part of people's life and sleep quality can tell a great deal about brain health. Research has found a strong link between abnormal sleep and brain disorders such as brain tumors[1, 2]. Therefore, people nowadays care increasingly more about their personal sleep quality, hence the development of several kinds of wearable devices to track sleep data[3, 4]. Studying sleep data used to deal with big machines, many different sensors, probes, and time-consuming manual sleep staging processes[5]. These old, complicated study methods not only carry the risks of human errors but also the high chances of inaccurate input data due to uncomfortable sleep because of attached devices on human subjects. With the development of technology, collecting sleep data is becoming easier than ever. The new compact wearable devices can be electroencephalography (EEG) devices that track the electrical activity of the brain; or they may be smart fitness watches such as Apple watches that gather heart rate, movement, steps, and other health and fitness data.

We should care about using wearable devices' data because these devices are becoming quite popular, so their data is more readily available and easier to get. However, the raw data does not yield sleep stage labels without being processed by certain algorithms. With the help of machine learning algorithms, we can categorize the sleep stages from smart wearable devices' data. This will lead to better utilization of this huge source of data available for sleep studies and easier sleep analysis. As a result, brain diseases linked to sleep disorders can be prevented and cured more efficiently and quickly.

Therefore, the primary objective of this project was to replicate the achievement of the work of classifying sleep stages (Wake/NREM/REM) using Apple Watch health data by Walch et al. Moreover, we wanted to make improvements on that work with more classes classified (Wake/N1/N2/N3/REM and/or Wake/N1+N2/N3/REM) as well as experiment different algorithms for different results.

Method

In the study, the subjects were given Apple Watches to monitor heart rate, acceleration, and circadian clocks. During an 8-hour night sleep, the subjects wore the provided Apple Watches while also underwent the golden-standard Polysomnography (PSG) monitoring. The PSG data were then manually scored into sleep stages by professional sleep technicians.

Our approach is to take the Apple Watch data and process it into a set of features, which will be then aligned and combined with the labels/target values obtained from PSG sleep score. The features and their corresponding labels will be used to train a classifier that can take that Apple Watch data and predict for each time interval what stage of sleep the subject was in. There are several distinctive features that will result from this data set, and each require various levels of processing to use for training. All data was cropped to figure out the latest start time, and earliest end time for each feature so that the results for each feature are consistent.

Heart Rate

The raw heart rate data from the Apple Watch has data in beats per minute (bpm) and is sampled every few seconds. After we interpolated that data to have one record for every second, a gaussian filter was applied to smooth the data. Additionally, all heart rate was normalized between subjects using the absolute difference between the heart rate each second and the mean heart rate over the entire sleep period. The final feature used is the standard deviation of the heart rate values.

Motion

The data for motion comes in from the Apple Watch in the form acceleration data in the x, y, and z directions in g, along with a timestamp. Like the heart rate data, interpolation was used to create a record for every second. Then an algorithm used to calculate activity counts based on the acceleration data. Gaussian filtering was again used to smooth data to create the file output of this feature.

Circadian

To calculate the circadian clock of the subject, we needed to substitute the typically used metric of light with steps, as the Apple Watch is not equipped with a useable light sensor. Thresholds for how many steps were needed for the subject to be considered awake were determined individually for each subject. If the subject was above that threshold for any time interval, the light levels were inferred based on the time of day to supply the necessary inputs to calculate the circadian clock.

Clock

This feature is used to consider the changes in the subjects' circadian rhythm over the course of the night. In this case, it was simply a cosine curve which started at the beginning of the sleep period and ended the end of the sleep period.

Polysomnography (PSG)

PSG data shows which sleep stage the subject was in, and this case, serves as our label for the data. The initial data uses the following integers to identify each sleep stage:

Stage	Wake	N1	N2	N3	N4	REM
Value	0	1	2	3	4	5

Figure 1. Sleep Stage integer mapping

We train the models using different features sets such as motion only, heart rate only, combination of motion and heart rate, or combination of motion, heart rate, and clock to explore which feature combinations are most useful and make improvements in model performance.

For training the algorithms, we used Monte Carlo cross-validation. The data is randomly split 10 times into a training set and a testing set, with a 70/30 training testing split ratio for subjects. Models will be trained on the data from the Apple Watch dataset, and the models developed there will then be used against the Multi-Ethnic Study of Atherosclerosis (MESA) dataset[6, 7] to see their performance on a different body of data.

The study we are basing our project on trained their models to perform two kinds of sleep stage categorization. The first is awake/asleep, so in terms of the PSG data, if the subject is 0 or 1-5. Their second was awake/NREM/REM, which categorized if the subject was awake (0), in REM (5) or in NREM (1-3).

As shown in Figure 1, formerly, non-REM sleep stages were divided into 4 stages N1, N2, N3, and N4. However, the current practice combined N3 and N4 into one stage N3, which is the deepest sleep stage. Hence, for this project, we are developing classifiers for 5 sleep stages - Wake (PSG value = 0) / N1(PSG value = 1) / N2 (PSG value = 2) / N3 (PSG value = 3 or 4) / REM (PSG value = 5). We also experimented with 4-stage classification Wake/N1-N2/N3/REM to see if it yields better performance.

We were able to replicate the initial study's metrics of accuracy and ROC curves but for each of the different sleep stages.

Experimental Results

By having a successful preprocessing pipeline created, we can take all the input data and generate clean and usable features for model training. This is shown in the below figure, which plots the results for each feature for one subject.

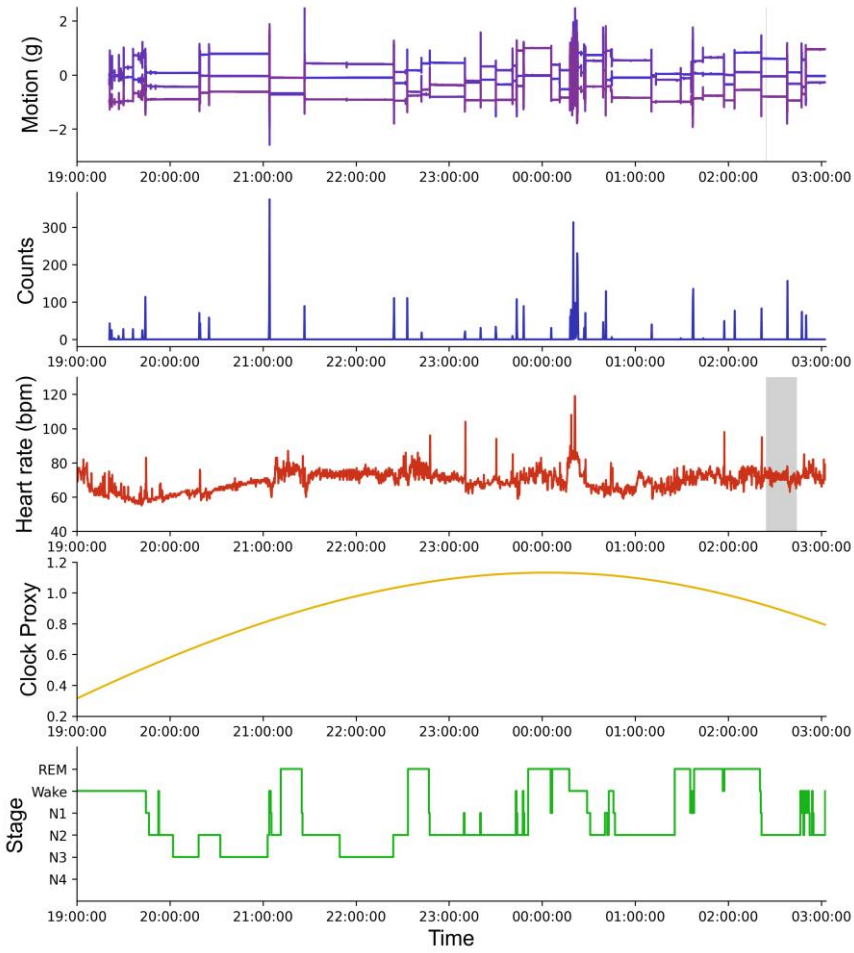


Figure 2. Feature visualization for Subject 1360686

Initially, we planned to perform ETL in an EMR (Elastic MapReduce) or Spark environment, but those frameworks were inappropriate for the code base and new directions we had to take. At present, we are using local workstations with Python (with Pandas, Scikit, SciPy libraries). The current data being used is at manageable enough size that local machines can successfully process it.

Initial result

Once we were able to successfully perform the necessary work to create usable features for training our models, we ran through some initial testing of converting the models to classify each of the five sleep stages: Wake, N1, N2, N3 and REM. While we plan to experiment with other different models and further hyper parameterization, as well as a four-stage classification of Wake, N1, N2/N3, and REM, we plan to use these initial results to form a baseline for our performance analysis.

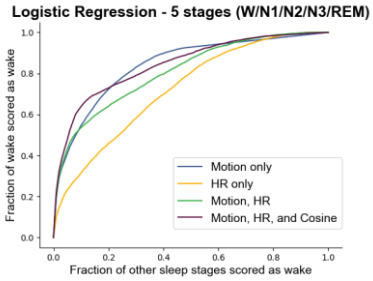
We have tested with three different machine learning algorithms, Logistic Regression, Random Forest, and K-Nearest Neighbors. ROC curve is typically used for binary classification to study the output of a classifier. Since our classifier is multi-class, we binarized the output and used one-versus-rest ROC. Shown below are the ROC curves which were generated by plotting the percentages of records incorrectly categorized as a particular stage in an epoch along the X-

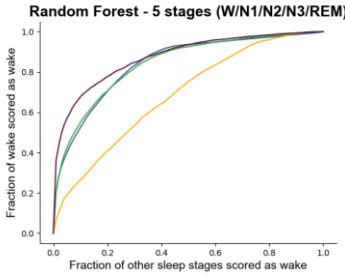
axis and the percentage of records correctly categorized as a specific stage in an epoch along the Y-axis. As this is a five-stage classifier, there are five charts shown to demonstrate the performance of the three models.

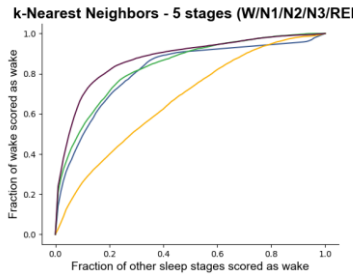
The closer a ROC curve is to the top left of the chart, the better the model, so when looking at the performance of each of these stages, the model does the best when classifying the Wake stage. The findings are consistently trending in the correct direction, with correct classifications showing as in 70%-80% range for each of the models, although K-Nearest Neighbors performs the best.

The model does start to struggle when it comes to analyzing the rest of the stages, however, with worse results. The main takeaways from this baseline are that the combination of Motion, HR and Cosine data typically has the best performance compared to other combinations of features. The performances of the different models at each stage are similar, but K-Nearest Neighbors still performs better, which suggests that clustering algorithms may be effective for these classifications. And while work will continue to be done to improve the performance of Logistic Regression and Random Forest, investigating K-Nearest Neighbors and other similar clustering algorithms will be a high priority.

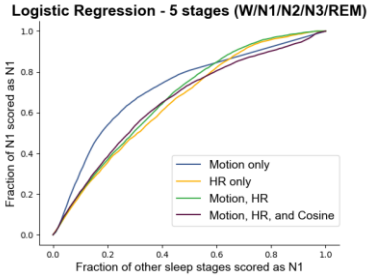
Wake sleep stage

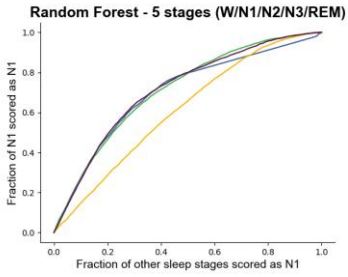


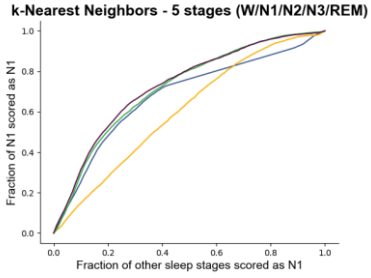




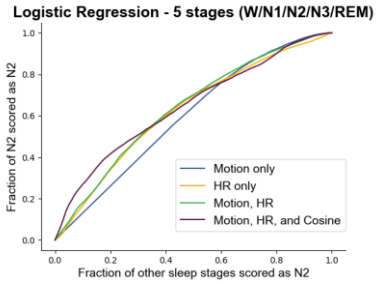
N1 sleep stage

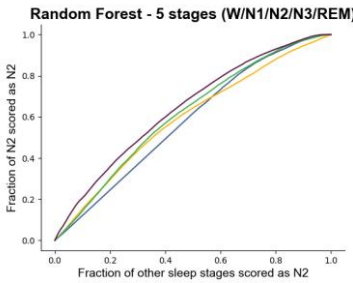


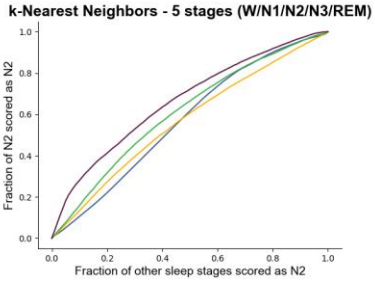




N2 sleep stage







N3 sleep stage

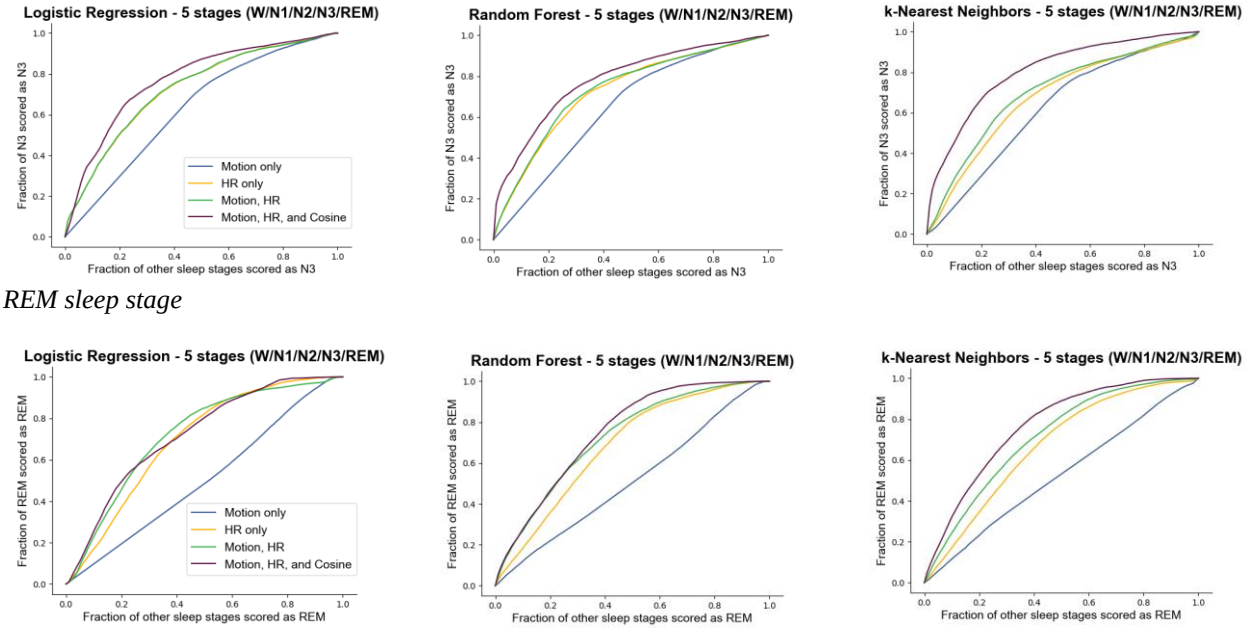
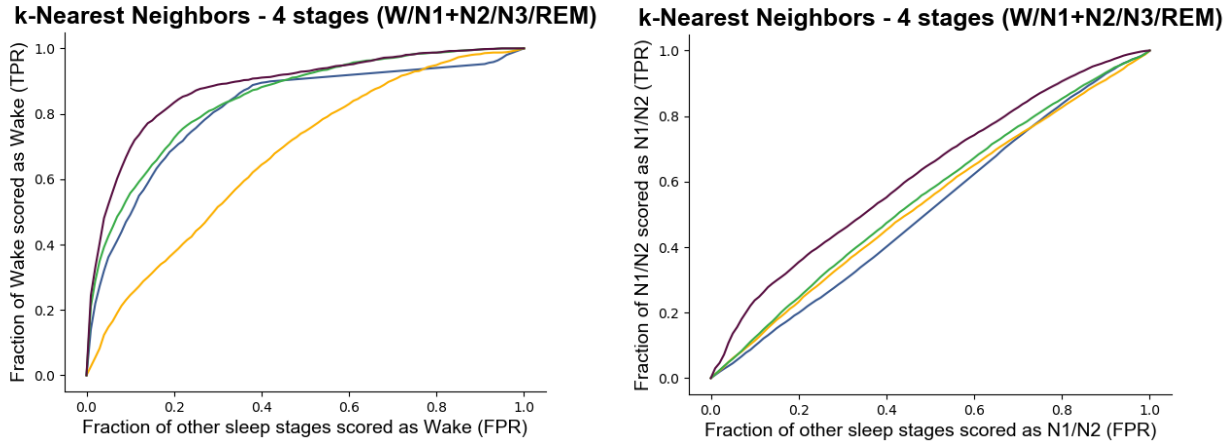


Figure 3. The above charts show one-versus-rest ROC curves for Logistic Regression, Random Forest, and K-Nearest Neighbors algorithms for a five-stage sleep classifier.

Further explorations

To improve performance of sleep stage identification, we tried adding code to classify sleep into 4 stages (Wake/N1+N2/N3/N5) instead of 5. Since N1 and N2 are both light sleep stages, we suspect they are more similar compared to other stages. However, the ROC curve for N1+N2 stage of the 4-stage classification performs worse than N1 and N2 ROC curves in the 5-stage classification, so we decided to focus on 5-stage classification.



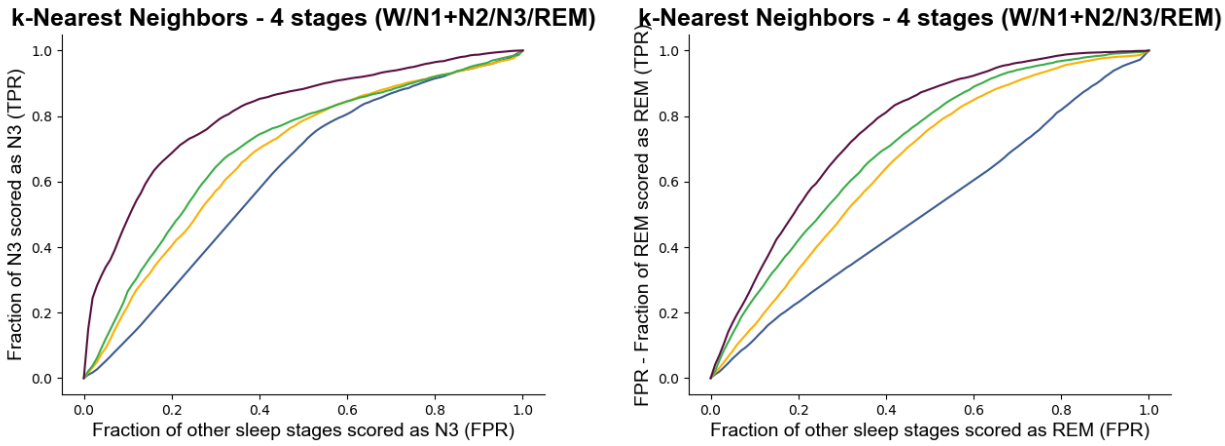


Figure 4. The above charts show one-versus-rest ROC curves for K-Nearest Neighbors algorithms for a four-stage sleep classifier.

In addition, we also trained additional algorithms. First, we added AdaBoost and the performance is not as good as k-NN and Random Forest. Secondly, we trained Multi-Layer Perceptron (MLP) Neural Network since it was the best performing algorithm in the original paper. Although we initially had problem with the neural net classifier taking too long to run, we did some adjustments to the code such as reducing the number of data splits. As a result, MLP classifier runs in an acceptable amount of time but still takes 10 times longer than other classifiers. MLP Neural Network model produces slightly better performance than k-NN model with similar trends for each sleep stage. The original paper's algorithm was implemented with Scikit, which has limited to support for other neural networks, and we were not able to integrate other types of neural networks.

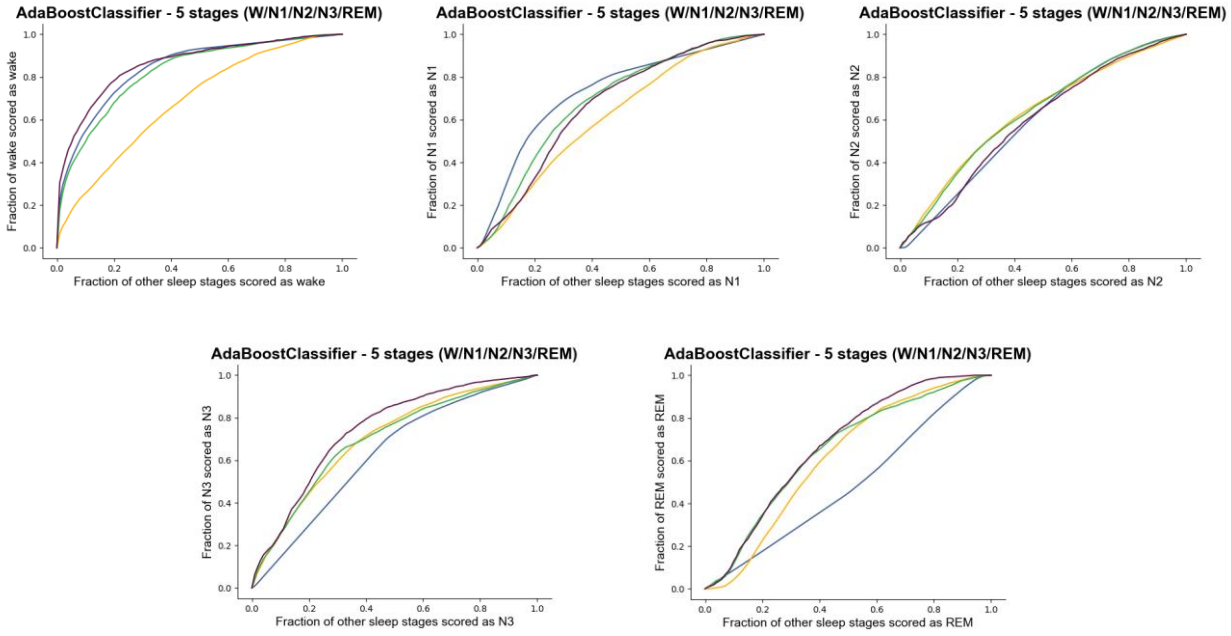


Figure 5. The above charts show one-versus-rest ROC curves for AdaBoost algorithm for a five-stage sleep classifier.

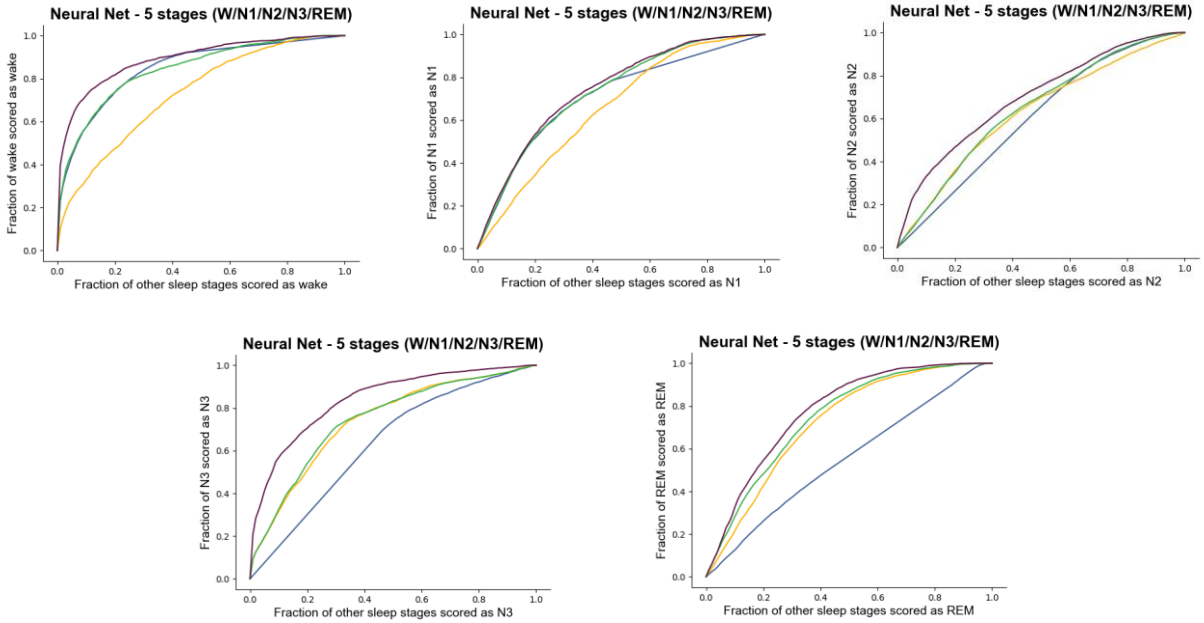


Figure 6. The above charts show one-versus-rest ROC curves for MLP Neural Network algorithm for a five-stage sleep classifier.

Discussion

While we intended to fully replicate the previous study first, we have faced a couple of problems. Firstly, in the original paper, the Apple watch data is used primarily for training and validating the model. The model is tested against the Multi-Ethnic Study of Atherosclerosis (MESA) dataset. We mistakenly assumed it would be easy to request, but it appears our access requests are taking a little longer than we expected. Secondly, replicating the initial study's results has been a challenge mostly due to figuring out performance bottlenecks on local machines. Results for one patient take at least a few hours to generate. Third, we originally planned to rewrite most of their ETL code to easily divide the tasks among group members. However, once we studied the code, we decided to work within the existing code base and run the code the whole way through on each of our local machines, as this gives us a more thorough idea of the entire implementation.

Given our timeframe, we had to scale back some of our ambitions. Particularly, the MESA data set is large and running the entire data set is unlikely to work on our current hardware setups (local workstations). We attempted to a subset of the MESA data and were able to replicate the results of the original work with only 3 classes (Wake/NREM/REM). However, we had some difficulties trying to classify 4 or 5 classes on the MESA data.

Therefore, we had to settle for the Apple watch dataset as our main dataset. Our goal shifted to adding sleep stages against the Apple watch dataset first, and if that was successful, move onto the MESA data. We also focused on improving Logistic Regression, Random Forest, and KNN algorithms. We tried different tweaks to those algorithms but could not improve performance compared to what we originally got.

While the original paper's code classified the sleep stages in 2 classes (Sleep/Wake) and 3 classes (Wake/NREM/REM), we added code to classify sleep stages into 4 and 5 classes.

Initially, we try to classify sleep data into 5 stages Wake/ N1/N2/N3/ REM because this is the most common way sleep researchers classify sleep. After trying multiple algorithms and our models show good result for wake, REM, and N3 class. N1 and N2 were not classified as well as other stages, especially N2.

To improve the performance, we tried different things such as tuning hyperparameter, training additional machine learning algorithms including AdaBoost and MLP, and experimenting with 4-stage classification. Among those

further experiments, only the MLP Neural Network model produces slightly better performance than initial k-NN model but still struggles with N2-stage classification.

While trying to improve ROC performance of N2 classification, we investigated the graph of feature data for each subject and noticed two things. First, the motion and heart rate curves tend to be flat during N2 stage, which means there are few or none motions, and heart rate are steady and slow during this stage. Second, there are a lot more samples for N2 than other stages, so there is a class imbalance problem. We suspect those two reasons might contribute to high False Positive Rate for N2 since Motion is the highest performing feature set for classifiers. We did some domain knowledge research and find out that the body temperature tends to drop[11] when entering N2 sleep stage, so if we have another feature set like body temperature recording, it might help classify N2 more accurately. Unfortunately, this feature is not available in the dataset we are using since the subjects were wearing Apple Watches, which do not have body temperature sensors. Future study setups might consider using other smart watches that have body temperate sensors like Fitbit. Another feature that might be also helpful is noise recording during sleep to detect snoring pattern.

Despite the above difficulties, we have made a critical improvement and expansion to original paper when our models show good performance for N3-stage identification. N3 is known as deep sleep or restorative sleep stage, which is very important for body recovery, growth, immune system boosting and contributes to insightful thinking, creativity, and memory[11]. Therefore, being able to identify correctly N3 sleep stage might help sleep study tremendously in measuring sleep quality.

Conclusion

The algorithms used in the original work yielded acceptable accuracy but left room for improvement. We successfully tried applying different algorithms other than Random Forest, Logistic Regression, and K-Nearest Neighbors, Neural Net, such as AdaBoost.

Furthermore, after successfully replicating and improving the models, we were able to increase the number of classes for classification. We got some good results for five sleep stage classification (Wake/N1/N2/N3/REM). We also built a classification with four sleep stages (Wake/N1+N2/N3/REM). Our models yield decent results but still have room for improvement.

With the development of technology and big data, various topics in health care are being studied more efficiently. Sleep care is an important part of well-being; hence many new technical devices have been developed to track sleep as well as well-being activities. Researching the data gathered by these devices will contribute to this trendy way of studying health care. We hope this work produced a well-performing classifier and obtain more sleep stages labeled. We also hope that this work can be improved even further and help encouraging more supports and drawing more attention to this topic. As a result, more detailed and accurate sleep information will be available and helpful in diagnosing and treating sleep disorders as well as improving sleep quality.

Team Contributions

Throughout every step of the project, our team members supported each other and worked on different tasks together where we saw fit with our skill set. In general, the contributions were split equally between team members.

References

1. Tsinalis O, Matthews PM, Guo Y, Zafeiriou S. Automatic sleep stage scoring with single-channel EEG using convolutional neural networks [Internet]. arXiv [stat.ML]. 2016. Available from: http://arxiv.org/abs/1610.01683
2. Wulff K, Gatti S, Wettstein JG, Foster RG. Sleep and circadian rhythm disruption in psychiatric and neurodegenerative disease. Nat Rev Neurosci. 2010;11(8):589–99.
3. Samuel N, So E, Djuric U, Diamandis P. Consumer-grade electroencephalography devices as potential tools for early detection of brain tumors. BMC Med. 2021;19(1):16.
4. Walch, O. (2019). Motion and heart rate from a wrist-worn wearable and labeled sleep from polysomnography (version 1.0.0). PhysioNet. https://doi.org/10.13026/hmhs-py35.
5. Zhao M, Yue S, Katabi D, Jaakkola TS, Bianchi MT. Learning sleep stages from radio signals: A conditional adversarial architecture. In 2017. p. 4100–4109,.
6. Zhang GQ, Cui L, Mueller R, Tao S, Kim M, Rueschman M, Mariani S, Mobley D, Redline S. The National Sleep Research Resource: towards a sleep data commons. J Am Med Inform Assoc. 2018 Oct 1;25(10):1351-1358. doi: 10.1093/jamia/ocy064. PMID: 29860441; PMCID: PMC6188513.
7. Chen X, Wang R, Zee P, Lutsey PL, Javaheri S, Alcántara C, Jackson CL, Williams MA, Redline S. Racial/Ethnic Differences in Sleep Disturbances: The Multi-Ethnic Study of Atherosclerosis (MESA). Sleep. 2015 Jun 1;38(6):877-88. doi: 10.5665/sleep.4732. PMID: 25409106; PMCID: PMC4434554.
8. Krystle Minkoff, 2016. "Sleep: The Evolution of Sleep Medicine in Neurology (Part One)," Open Access Journal of Neurology & Neurosurgery, Juniper Publishers Inc., vol. 2(1), pages 7-8, December.
9. Olivia Walch, Yitong Huang, Daniel Forger, Cathy Goldstein, Sleep stage prediction with raw acceleration and photoplethysmography heart rate data derived from a consumer wearable device, Sleep, Volume 42, Issue 12, December 2019, zsz180, https://doi.org/10.1093/sleep/zsz180
10. Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation [Online]. 101 (23), pp. e215–e220.
11. Stages of Sleep [Internet]. Sleepfoundation.org. 2020 [cited 2021 May 2]. Available from: https://www.sleepfoundation.org/how-sleep-works/stages-of-sleep