Sleep stage estimation using RNNs with physiological data from wearable devices

January 3, 2022

1 Introduction

On average an individual spends about one-third of their life in sleep. Sleep is vital for effective functioning and health in humans. Getting adequate sleep every day can help one maintain/improve mental, physical health, and overall quality of life. On the other hand, lack of sleep can lead to problems such as difficulty in decision making, emotional imbalance; persistent sleep deprivation is associated with a higher risk of chronic health problems. In a nutshell, good sleep is important for a healthy lifestyle in humans.

The last decade has seen tremendous growth in the popularity of consumer wearable devices. At the same time, there is also an increasing emphasis on tracking personal health metrics. Sleep tracking is one such aspect with implications on overall health and lifestyle quality. Currently, there are many products in the consumer wearables market aimed at measuring sleep quality and efficiency [Shelgikar et al., 2016]. These devices typically collect numerous data from multiple sensors such as heart rates, movements, body temperature, among others. Subsequently, various factors such as sleep stages, sleeping timings, latency are considered while estimating the overall sleep quality and efficiency for end consumers.

In most wearable technologies, the underlying algorithm is undisclosed and the access to all the metadata considered by the algorithm is restricted. The reliability of these metrics given by the devices is also questionable in many cases [Kolla et al., 2016]. Therefore, this project aims to study the feasibility of estimating sleep stages based on the observations made by wearable devices. In particular, we consider the problem of estimating sleep stages based on heart rate and movement metrics, commonly measured with many wearable devices.

We consider the open-source data available on `physionet.org` [1] reporting personal activity, physiological observations, and sleep stage measurements. We closely follow the accompanying approach proposed by the authors responsible for open-sourcing the data [Walch et al., 2019], and empirically evaluate different algorithms to address the sleep stage estimation problem. The prior work of Walch et al. [2019] explored sleep stage classification based on epochs of 30 seconds in isolation. This previous work considered classification techniques such as random forests, k-nearest neighbors, logistic regression treating every epoch independently over time. This independence over time assumption can be restrictive as

[1] https://physionet.org/content/sleep-accel/1.0.0/

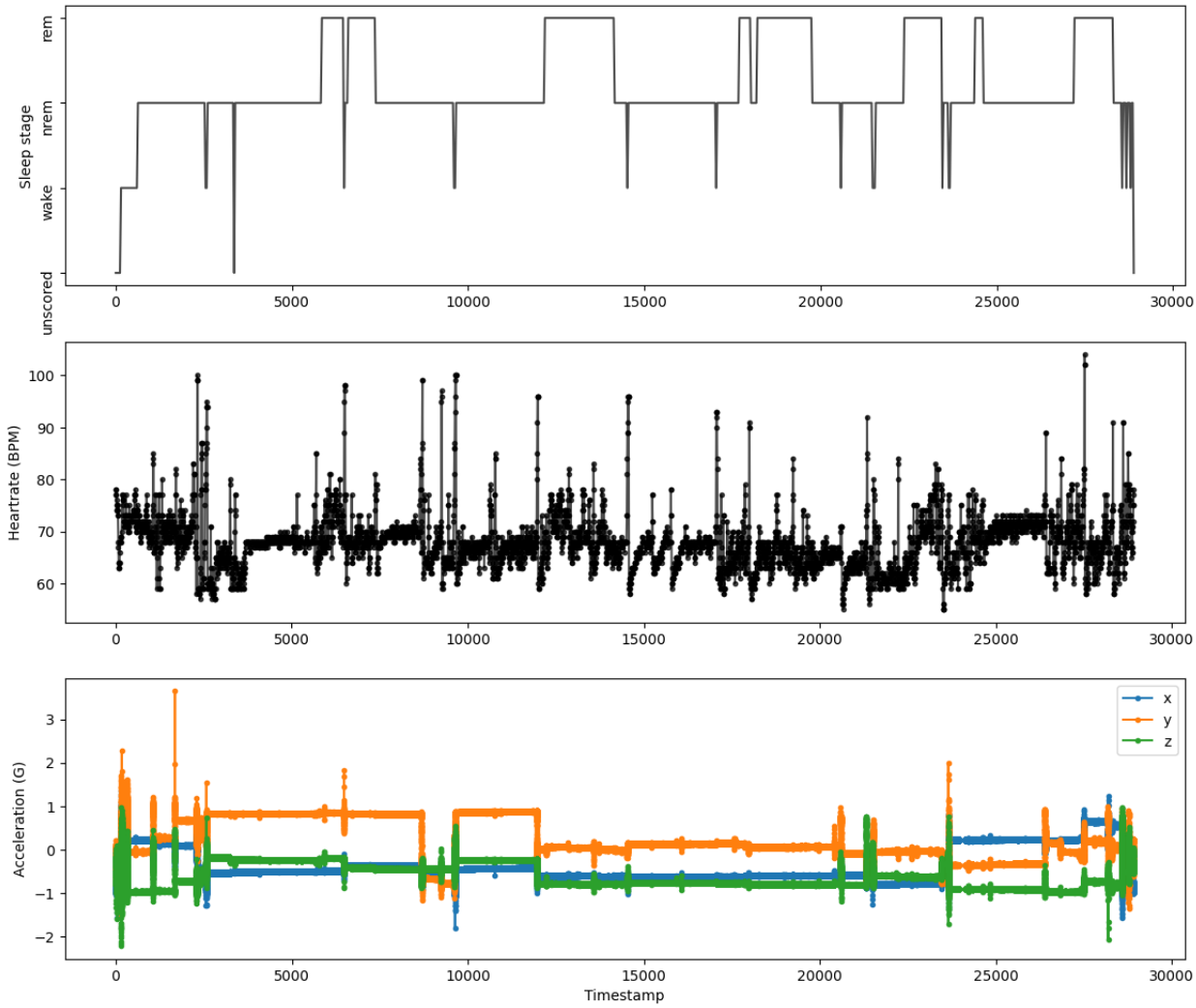

Figure 1: Visualization of the raw data for subject 8000685. The dataset consists of sleep stage labels tagged at 30 second epochs. In addition, the heart rate and acceleration parameters are measured at irregular time intervals.

it ignores temporal correlations in the data. In the current work, we considered state-space methods using RNNs that can model the temporal structure of the data. The more recent works on the sleep estimation problem also considered deep neural network based approaches [Biswal et al., 2018, Radha et al., 2019, Haghayegh et al., 2021], however on datasets based on heart rate variability measurements that are different than the one under study . With the proposed RNN model, we observed around 10% increase in classification performance. Empirically, we observed that we can estimate different sleep stages based on heart rate and motion measurements with around 70% accuracy.

2 Problem formulation

Normal sleep in a healthy individual comprises two different stages: rapid eye movement (REM), and non-rapid eye movement (NREM). One sleep episode typically consists of alternating cycles through the REM and NREM sleep stages, along with possible wakefulness states in between. The NREM stage can be further divided into four stages N1, N2, N3, and N4, with N1 & N2 typically considered "light sleep" and N3 & N4 "deep sleep" states. Characterization of good sleep in an individual requires an understanding of personal sleep stage progression. The currently established gold standard for sleep measurement is the polysomnogram (PSG). However, quantifying sleep with PSG is an expensive process, as it requires measuring multiple physiological parameters in a sleep laboratory [Berry et al., 2015]. More accessible ways of measuring sleep can be helpful for wider utilization of sleep medicine, which can bring health and lifestyle improvements for many people.

Given the difficulty in large-scale usage of polysomnography for personal sleep tracking, estimating sleep using other consumer wearable devices can be a convenient solution. Such wearable devices are commonly affordable, track multiple personal biological signals, and adopted by the wider population over recent years. In the context of sleep tracking, measurements of motion with accelerometers, heart rate, and respiratory signals with photoplethysmography (PPG), body temperature measurements can be useful. Such parameter tracking is available in many wearable devices today. Such variables measured can be thought to have a predictive relationship with the sleep stages. In this project, we consider the problem of estimating sleep stages using data regarding heart rate and motion variables with different machine learning predictive modeling techniques.

Sleep data can be thought of as a dynamic process, with temporal structure in place. In particular, there are specific patterns generated at different sleep stages regarding the measurable variables, which can be utilized for predictive modeling. In specific, we consider heart rate and motion as measurables that can have explicit patterns in NREM/REM/ Wake states in a typical sleep episode. Hence, we can treat the sleep stage estimation problem as a classification problem with observed measurements as input variables and gold-standard sleep stage labels from PSG as target variables. Given that we can build such a model with good accuracy, we can utilize such a model for estimating sleep stages for a general population using wearable devices.

3 Dataset description

3.1 Data generation

The dataset used in this work consists of the data recorded for 31 subjects at the University of Michigan as part of the work by Walch et al. [2019]. The sleeping data for all the subjects were recorded with PSG over a course of one night with an opportunity to sleep for 8 hours. The subjects were also asked to wear Apple Watch (Apple Inc.) while undergoing PSG recording. The Apple watch was used to record motion data with MEMS accelerometers and heart rate data with photoplethysmography (PPG). Walch et al. [2019] also provide an

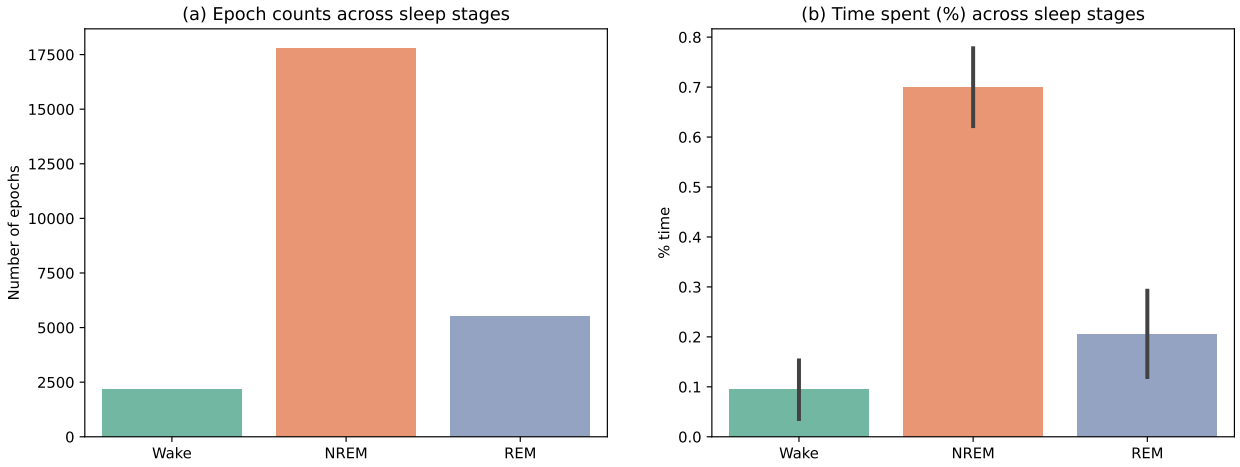

Figure 2: The modeling dataset consists of three classification labels : NREM, REM and Wake. Around 70% of epochs or observations are od NREM stage, followed by around 20% of REM, and around 10% of Wake states as shown in subplot (a). The CI in subplot (b) shows the variation across subjects : the subject-level sleep distribution is similar to the population-level distribution. This observation is consistent with the general understanding of sleep [Carskadon et al., 2005].

open-source codebase for accessing heart rate and accelerometer data from Apple watch [2].

3.2 Raw dataset

The raw dataset consists of true sleep stages over time as recorded by PSG along with the motion and heart rate data over time, for the duration of one night.

Sleep stage data with PSG The dataset consists of sleep stage measurements for 31 subjects using PSG. Each subject can have a possibly different sequence length based on their sleep duration on the night of study. These sleep stages consist of 6 labels: REM, N1, N2, N3, N4, Wake, and Unscored. The PSG labels were recorded by considering measurements of multiple biological signals using electroencephalogram (EEG), electrooculogram (EOG), electrocardiogram (ECG), among others following the technical specifications of the American Academy of Sleep Medicine (AASM) [Berry et al., 2015] in the laboratory. We refer readers to Walch et al. [2019] for further information on the PSG recording process.

Heart rate and motion data In addition to the sleep stage labels, the heart rate data in beats per minute (BPM) and motion data in units of g ($9.8m/s^2$) using Apple watch. The motion data consists of accelerometer reading in x, y, and z directions. Both heart rate and acceleration data are measured at irregular time intervals as shown in Figure 1.

[2] https://github.com/ojwalch/sleep_accel

3.3 Modeling dataset

Classification targets The raw dataset consists of six sleep stage labels (REM, N1-N4, Wake) along with an additional "Unscored" label. These PSG labels are generated for every 30 second time interval called "epochs". The raw dataset was filtered to exclude the "Unscored" epochs while generating the modeling dataset. In addition, the four sleep stages N1-N4 were combined to form the NREM stage for classification modeling target with three classes: REM, NREM, and Wake stages.

The population-level histogram or marginal distribution of the sleep stage labels are shown in Figure 2. The distribution is not uniform across different labels as shown in the figure. NREM is the majority class covering around 70% epochs, followed by REM of 20% and Wake state around 10%. This indicates that the class imbalance needs to be carefully considered while building the classification model.

Classification inputs While the PSG labels are recorded in regular 30-second epochs, the motion and heart rate (HR) data are recorded from Apple watch at irregular time intervals. As a preprocessing step, these irregular HR and motion data were interpolated to one-second intervals for every subject. This resulted in each sleep epoch having 30 measurements of HR and motion data respectively. The 30 measurements for every epoch were further aggregated to produce different modeling input features as described below.

The HR readings were aggregated to two different features (1) average HR feature, and (2) HR variation feature. Average HR corresponds to the mean HR value for the 30-second epoch interval. The HR variational corresponds to the standard deviation of the heart rate differentials generated by convolving the interpolated HR signal with 120 second and 600-second difference of Gaussian filters.

The motion readings were also aggregated to two different features (1) average motion feature, and (2) motion variation feature. These features are computed as mean and standard deviation values of total motion (defined as squared L2 norm) of interpolated accelerations in the 30-second epoch intervals.

The physiological variables such as HR and motion can be considered as observable responses of underlying sleep states in subjects. While such variables could help build a predictive model, we could also consider the external environmental influence on an individual's sleep. Such external variables can be considered under a 24-hour cycle referred to as the "circadian cycle" that can indicate sleep propensity in individuals. In this work, we follow the approach of Walch et al. [2019] and include an additional time-based feature that maps the clock-time to a cosine signal for individual sleep episodes as shown in Figure 3.

4 Methods

4.1 Baseline models

In the work accompanying the dataset under study, Walch et al. [2019] proposed classification models for sleep stage prediction using acceleration and heart rate measurements. The modeling formulation in this work treated every sleep epoch in isolation. Specifically, this prior work explored models such as random forests, logistic regression and provided an

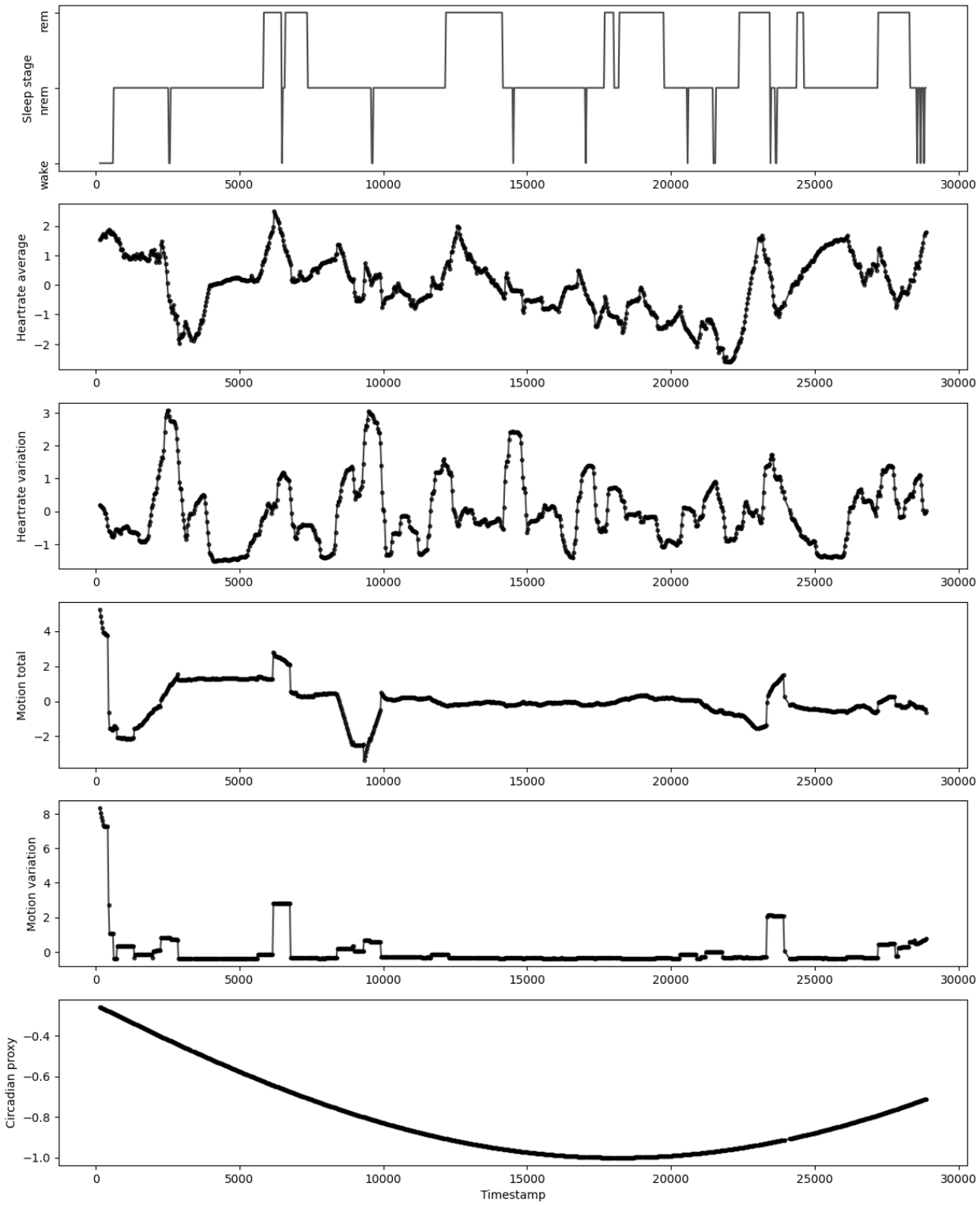

Figure 3: Visualization of the modeling features for subject 8000685. The modeling dataset consists of sleep stage labels at regular 30 second epochs. The input features consists of two heart rate related features: average heart rate and heart rate variation, two motion related features: average motion and motion variation, and a time-based circadian proxy feature.

open-source codebase [3]. In this work, we trained models similar to Walch et al. [2019] as baselines on the modeling dataset described in Section 3.3.

4.2 State-space model

One of the shortcomings of the baseline modeling approach is that they treat each epoch of sleep in isolation. This yields "blips" of sleep stages and ignores temporal correlations in the dataset [1]. As a new work, we considered state-space methods based on recurrent neural networks (RNNs) that can incorporate latent temporal structure into the predictive model. In terms of modeling framework, the current work is similar to Radha et al. [2019], Haghayegh et al. [2021] that utilize deep neural networks for sleep stage classification. The RNN model chosen consists of a GRU block [Cho et al., 2014] mapping observed classification inputs to latent dynamics space. The latent space dynamics are then mapped to a fully connected block with one hidden layer, followed by the softmax classification layer. The latent units in the model can incorporate historical information from the input sequence.

4.3 Training setup

The modeling dataset consists of five feature sequences and one classification target sequence for 31 subjects. The full dataset was split into 8 training/testing sets with 30% subjects selected for testing at random. For baseline classification models, we selected the best-performing parameters based on grid search using the training split. For RNN models, to avoid overfitting, the training dataset was further divided into train/validation splits randomly, and validation loss was used for early stopping. We used negative log-likelihood as the objective function for all models under consideration.

5 Results

We considered two baseline models logistic regression and random forests along with a new RNN model in this study. We trained models on 8 different training/testing splits independently and report the average values for different performance metrics in this section. Figure 4 compares the ground truth sequence to the label sequence predicted by different classification models for the test subject `8000685`.

The overall classification performance in terms of different metrics is reported in Table 1. As seen from the table, all the models learn to classify three different sleep stages at more than 65% accuracy score. The results for baseline classifiers are similar to those reported by Walch et al. [2019] on the dataset. The logistic regression classifier performs the best in terms of overall accuracy. However, the overall accuracy score might not be the best metric of choice for assessing the classification performance in the case of the imbalanced dataset as the metric is biased towards the majority label in terms of the number of observations. Figure 7 shows distribution of classification agreements across sleep stages for different classifiers. From the figure, we can see that all three sleep stages are predicted with almost similar accuracy in the case of RNN and random forest models, whereas the logistic regression is more biased

[3] https://github.com/ojwalch/sleep_classifiers/

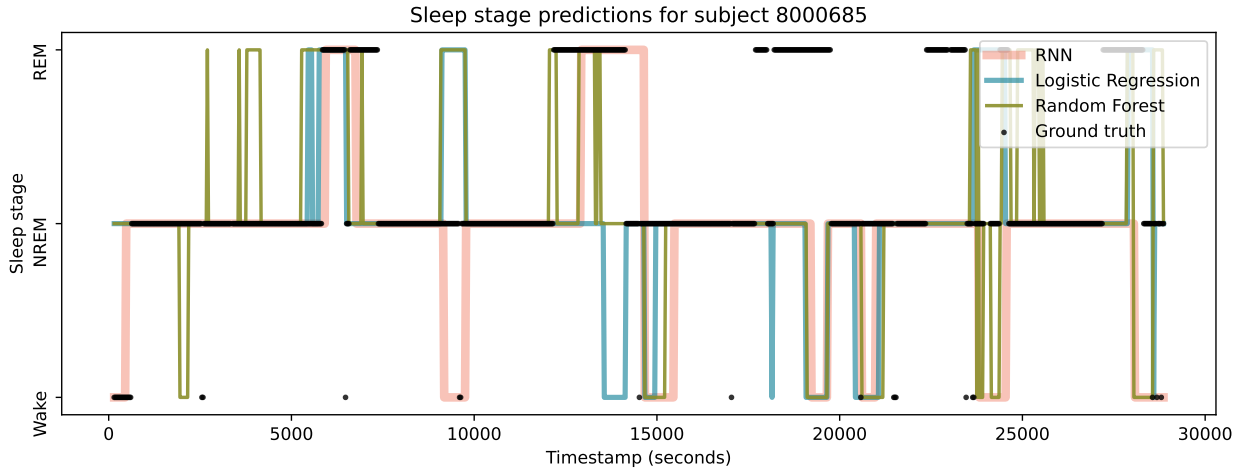

Figure 4: Visualization of the model predictions for subject 8000685 for three different models RNN classifier, logistic regression and random forest as compared to the ground truth. All the models give misclassified assignments without a distinct misclassification pattern.

Table 1: Classification metrics for test subjects. The best values bolded. (↑): higher is better, (↓) lower is better

Method	Acuracy (↑)	Balanced acuracy (↑)	ROC AUC (↑)	Cohen κ (↑)
RNN	0.676	**0.685**	**0.851**	**0.418**
Logistic Regression	**0.718**	0.573	0.786	0.359
Random Forest	0.666	0.593	0.796	0.342

towards the dominant NREM class. In addition, the confusion matrix (normalized by the number of observations in the dataset) reported in Figure 5 indicates that while the logistic regression model predicts the majority class NREM accurately, it fails to predict the REM and Wake classes accurately.

We can instead consider assessing the classification performance based on the balanced accuracy score, which computes the average recall obtained on each class. From Table 1, we can see that the RNN model gives a better-balanced accuracy score as compared to the other baseline methods. In addition, Figure 6 shows the one-vs-rest ROC scores for all three classification models. As shown in the figure, the RNN model performs better than the other baseline models, the result of which is also quantified in terms of area-under-the-curve (AUC) score reported in Table 1. In addition, we also included Cohen's kappa statistic [Cohen, 1960] that takes into account the potential for classification agreement to happen by chance. In terms of this static, the RNN model again outperforms the other baseline models as shown in Table 1.

In addition, we follow the multi-class sleep stage classification analysis of [Walch et al., 2019] and report the results in this work. This analysis applies two thresholds to the class probabilities predicted by the classification models. The first threshold was applied to keep the false positive rate for the Wake class at 60%. A second threshold was then applied to those epochs not scored as Wake under the first threshold. This second threshold was chosen

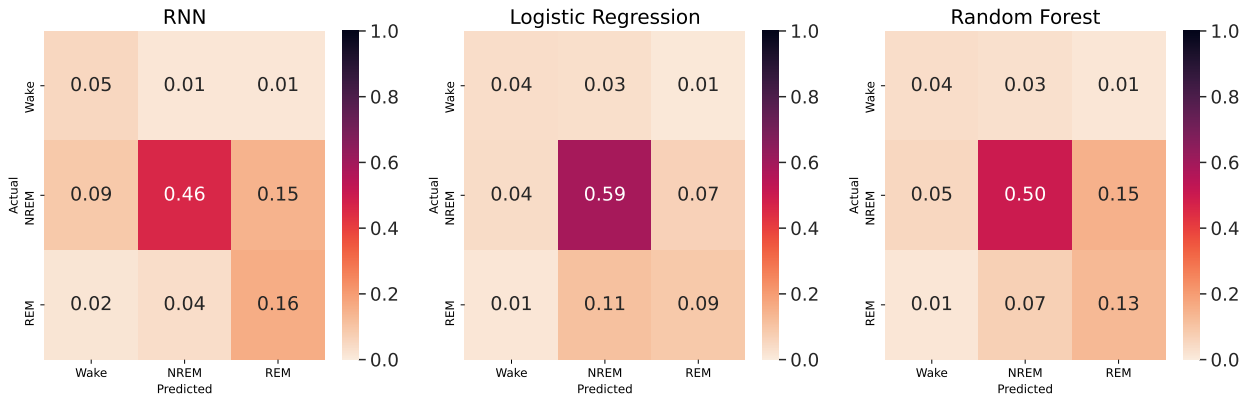

Figure 5: Confusion matrix (normalized by the number of observations) for different models.

Table 2: Classification metrics for test subjects based on binary search. The best values bolded. (↑): higher is better, (↓) lower is better

Method	Wake correct ↑	NREM correct ↑	REM correct ↑	Cohen κ ↑
RNN	0.600	**0.720**	**0.720**	**0.465**
Logistic Regression	0.600	0.633	0.631	0.303
Random Forest	0.600	0.610	0.609	0.333

to maintain the REM and NREM class accuracies as close to equal as possible. As noted by [Walch et al., 2019], this method of analyzing classification models requires the true labels for testing sets. Hence this analysis should be only considered for understanding the model performance when the ground truth is known. The quantitative results from this analysis are reported in Table 2. From the numbers, we can see that the classifier RNN model can accurately classify NREM and REM stages with 72% accuracy which is more than 10% improvement over other baseline methods. In terms of the Cohen κ statistic, the RNN again outperforms the baseline classifiers.

6 Conclusion and discussion

We considered the problem of estimating sleep stages based on physiological observations and a proxy feature for circadian rhythm. In particular, the study explores the possibility of identifying sleep stages based on motion and heart rate measurements from a wrist wearable device. While the previous methods on the dataset under study modeled each sleep epoch isolation, such modeling methodology ignores the temporal correlations in the sleep data. As an attempt at improving the sequence classification modeling, we considered a state-space method based on RNNs in this study. The results show that incorporating such temporal structure into the model can help in improving the sleep stage classification problem.

In this work, we considered a straightforward implementation of RNNs for modeling temporal dynamics in the sleep data. From a machine learning modeling perspective, there are multiple rooms for improvement. Within the current modeling framework, one could consider

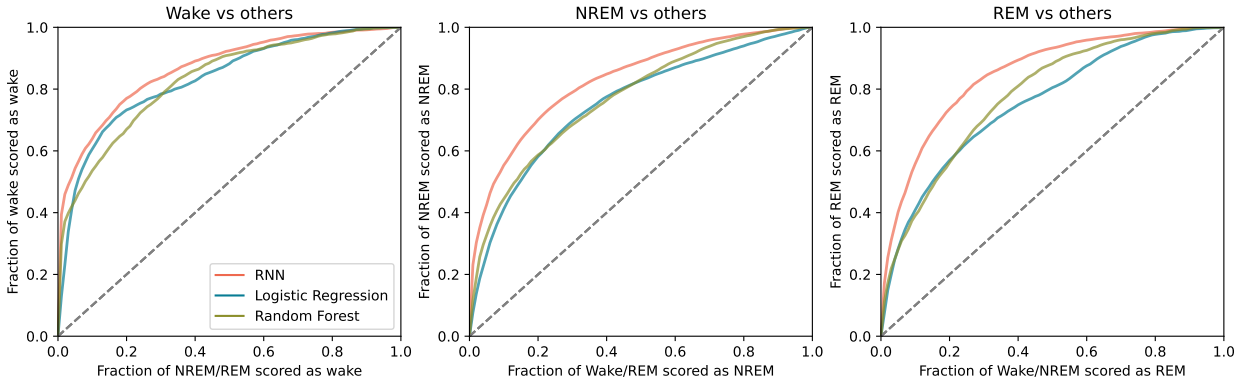

Figure 6: One vs rest ROC score comparison for sleep stage classification problem across different models. While the most prominent NREM state is classified almost equally well by all three models, RNN-based model performs better compared to the baselines on REM and Wake stages.

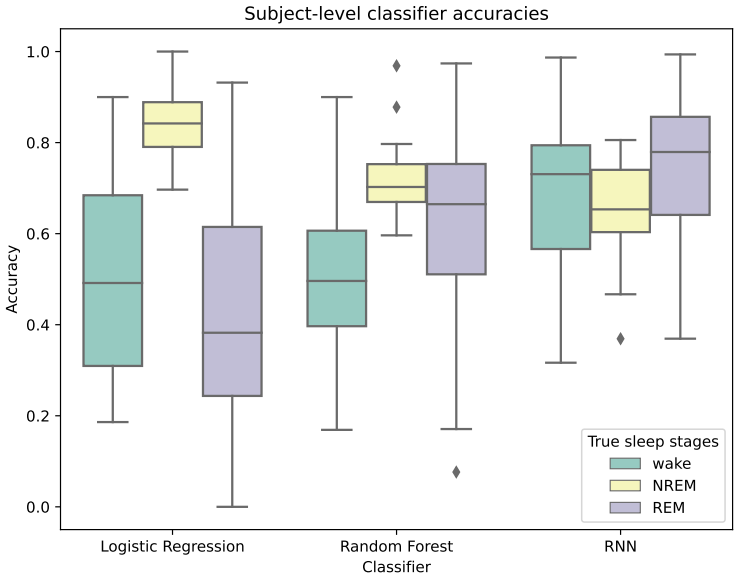

Figure 7: Comparison of the subject-level classification accuracies across different models. Logistic regression gives good accuracy on the NREM majority class, while gives poor accuracy on REM and Wake. In contrast, random forest and RNN classifiers provide almost equal accuracy on all three classes, with RNN performing best in terms of average accuracy.

deriving more informative features such as the circadian feature based on the ambulatory data as defined in Walch et al. [2019].

The dataset under study consists of raw measurements at irregular time intervals. In this work, we generated aggregated raw data to uniform time intervals (sleep stage epochs) by generating features corresponding to the first and second-order statistics (mean and variation respectively). However, the higher-order information could be lost because of this aggregation, which might be useful for predictive modeling. A possible future direction could be to develop models that can handle irregularly sampled time series data [Sun et al., 2020]. The dataset understudy had only one night of observation for a total of 31 subjects, which can be quite limited. Another line of extension could be to consider probabilistic approaches that could provide better generalization under the limited number of observations [Fortunato et al., 2017].

In this current work, we considered the utilization of heart rate and motion measurements for predicting sleep stages. The state transitions throughout a sleep episode are also known to generate variations in other observables such as body temperature and respiration rates among others. One could incorporate such features in the predictive modeling that are tracked in many consumer wearable devices. In addition, we considered a single predictive model at the population level with an assumption that the features capture sufficient individual-level information. Inclusion of additional relevant information such as demographic parameters could be essential for generalization to wider populations.

References

Richard B Berry, Rita Brooks, Charlene E Gamaldo, Susan M Harding, C Marcus, Bradley V 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*, 2015.

Siddharth Biswal, Haoqi Sun, Balaji Goparaju, M Brandon Westover, Jimeng Sun, and Matt T Bianchi. Expert-level sleep scoring with deep neural networks. *Journal of the American Medical Informatics Association*, 25(12):1643–1650, 2018.

Mary A Carskadon, William C Dement, et al. Normal human sleep: an overview. *Principles and practice of sleep medicine*, 4(1):13–23, 2005.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

Meire Fortunato, Charles Blundell, and Oriol Vinyals. Bayesian recurrent neural networks. *arXiv preprint arXiv:1704.02798*, 2017.

Shahab Haghayegh, Sepideh Khoshnevis, Michael H Smolensky, Kenneth R Diller, and Richard J Castriotta. Deep neural network sleep scoring using combined motion and heart rate variability data. *Sensors*, 21(1):25, 2021.

Bhanu Prakash Kolla, Subir Mansukhani, and Meghna P Mansukhani. Consumer sleep tracking devices: a review of mechanisms, validity and utility. *Expert review of medical devices*, 13(5):497–506, 2016.

Mustafa Radha, Pedro Fonseca, Arnaud Moreau, Marco Ross, Andreas Cerny, Peter Anderer, Xi Long, and Ronald M Aarts. Sleep stage classification from heart-rate variability using long short-term memory neural networks. *Scientific reports*, 9(1):1–11, 2019.

Anita Valanju Shelgikar, Patricia F Anderson, and Marc R Stephens. Sleep tracking, wearable technology, and opportunities for research and clinical care. *Chest*, 150(3):732–743, 2016.

Chenxi Sun, Shenda Hong, Moxian Song, and Hongyan Li. A review of deep learning methods for irregularly sampled medical time series data. *arXiv preprint arXiv:2010.12493*, 2020.

Olivia Walch, Yitong Huang, Daniel Forger, and Cathy Goldstein. Sleep stage prediction with raw acceleration and photoplethysmography heart rate data derived from a consumer wearable device. *Sleep*, 42(12), 08 2019.