Sleep cycle classification from Apple Watch data

Final project for the course CS-C4100 Digital Health and Human Behaviour

1 Introduction

When we fall asleep, our body cycles through two stages of sleep, namely the rapid eye movement (REM), and non-rapid eye movement (NREM) stages. The non-REM phase can be further divided into three stages, labeled as N1, N2, and N3 (Patel, 2022). During the night, we usually undergo four to six of these cycles, each lasting around 80 to 100 minutes (*Sleep Phases and Stages*, 2022).

These four stages of sleep (REM, N1-N3), characterized by the National Health, Lung and Blood Institute (NIH), are presented in the following Table 1.

Table 1: Description of Sleep Stages

REM	During this stage, rapid eye movement occurs, and the brain is highly active. REM phase is associated with vivid dreaming.
N1	The first stage of NREM sleep. Transition from wakefulness to sleep. Muscle activity decreases, and conscious awareness of the external environment diminishes.
N2	Slightly deeper stage of non-REM sleep, characterized by a decrease in heart rate and a further reduction in muscle activity. It represents a state of light sleep.
N3	Also known as deep sleep, N3 is the stage where the body and brain undergo significant restoration. This stage is crucial for physical recovery, growth, and overall well-being.

Studying sleep stages is crucial for effective diagnosis and treatment of sleep-related disorders, with traditional manual scoring being time-consuming and error-prone (Sharma et al., 2021). Hence, scoring sleep stages through machine learning techniques is essential to attain precise diagnoses.

Version December 11, 2023

By far, polysomnography (PSG) methods utilizing electroencephalogram (EEG) measurements are among the most frequently used ones for sleep stage scoring (Gaiduk et al., 2023).

The use of different classification methods for sleep stage identification has also been widely studied. The state-of-the-art methods, especially Random Forest (RF) classifiers, are proven to produce promising results with high accuracies (Boostani et al., 2017). These results are further backed by the study by Sharma et al. (2021), which employs a decision tree-based ensemble classifier model which achieves notably high accuracies on classifying sleep stages.

However, the current sleep scoring classification models rely heavily on PSG data, particularly electroencephalogram (EEG) measurements. The laboratory measurements are restricted for the most severe sleep disorders, which limits the broader applicability of these classification methods (Miettinen et al., 2018).

Thus, the objective and research question of this report is to find out whether it's possible and feasible to extract sleep stage information in a non-clinical setting using an consumer wearable device.

The classifier discussed later in the report achieved an overall accuracy of 69,5% in classifying sleep stages, with notable success in identifying NREM stages (f1-score 0.79). However, challenges were observed in accurately classifying REM phases, particularly in distinguishing between wake and REM stages.

2 Problem Formulation

This report aims to study the use of data gathered from a consumer wearable device (Apple Watch), to predict the sleep stages of individuals. Given the restricted access to clinical equipment, the study seeks to assess the viability of using readily available sensor data to classify sleep cycles within a real-world, non-clinical context.

The research problems are

1. Can machine learning models trained on data from consumer wearable devices achieve comparable sleep stage classification compared to tradition PSG methods?
2. What are the key opportunities and limitations of these classifications models?

The objective of the project is to construct useful features from the given dataset and to predict the sleep stages using a Gradient Boosting classifier. The results and findings will be then analysed and further compared to the ones mentioned in the literature and the original publication.

3 Dataset

The dataset, collected at the University of Michigan from June 2017 to March 2019 (Walch, 2020), contains motion (*g* forces in x, y, and z), heart rate (*bpm*), steps (*count*), and labeled sleep data (wake = 0, N1 = 1, N2 = 2, N3 = 3, REM = 5) across 31 subjects. The data was collected using Apple Watches during the period of the study.

The raw data (motion, heart rate, steps, labels) are presented as timestamped measurements stored in separate `.txt` files. The filename contains the measured variable and the subject's anonymized id. The timestamps are expressed as seconds (s) prior PSG start, i.e., a negative timestamp implies that the value was measured prior the start of PSG.

Before further feature engineering, the raw data was processed using the Python script supplied in the foundational work conducted by Walch (2020). As the study aims to predict the sleep stage during sleep, all datapoints prior the sleep are discarded.

The following subsections present the selected features and their corresponding transformations.

3.1 Heart rate

The heart rates were recorded using the photoplethysmogram sensor from the Apple Watch. The heart rate measures were saved in irregular intervals spanning from a few seconds to around 10 seconds. It was stored as a decimal numbers representing the beats-per-minute.

To use the heart rate as a feature, some proprocessing, namely interpolation, differencing, and normalisation, was applied.

Using the data preprocessing script by Walch (2020), the granular raw data was first interpolated to have a value every second, and further processed by convolving with a difference of two Gaussian filters. Finally, each individual's readings were normalized by dividing the features by the 90th percentile in the value of the first-order difference of the measurements. The heart rate feature before and after the transformation is presented in Figure 1.

3.2 Motion

Motion data was acquired from the Apple Watch at 50 hz acceleration readings for three axis, x, y, and z. The measurements were stored as decimal numbers representing the force g. The raw acceleration readings were subsequently transformed into an activity count feature by signal processing (band-pass Butterworh filter) using the script by Walch (2020). The raw acceleration readings and the activity count feature are presented in Figure 2.

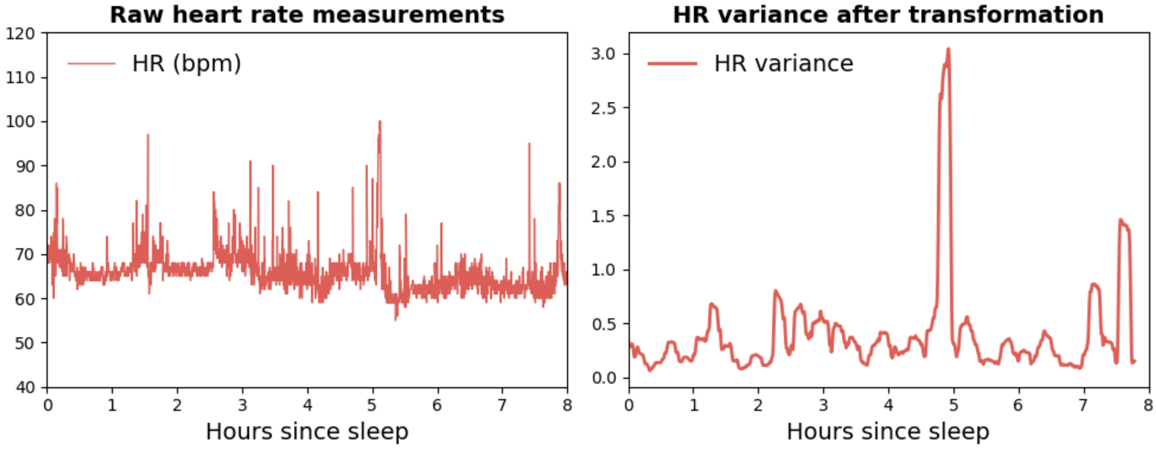

Figure 1: Heart rate measurements from the Apple Watch (left), and the transformed heart rate feature (right).

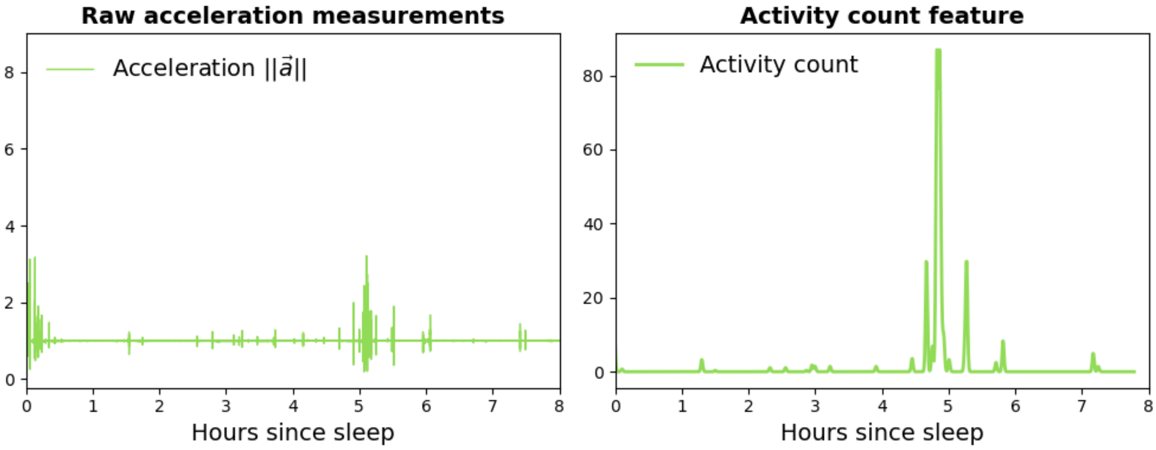

Figure 2: Raw acceleration measurements (left) and activity count feature after transformation (right).

3.3 Circadian phase

The circadian phase plays an important role in understanding the temporal patterns of physiological and behavioral changes. The circadian cycle refers to the 24-hour behavioural cycle, influencing various aspects of the daily life, including our sleeping patterns.

Following the method proposed by Walch (2020) shown in Figure 3, the circadian cycle can be modelled using a "proxy clock" determined both by a fixed cosine wave based on the time of the recording, and a mathematical model (Forger et al., 1999) for incorporating personalized, longitudinal information.

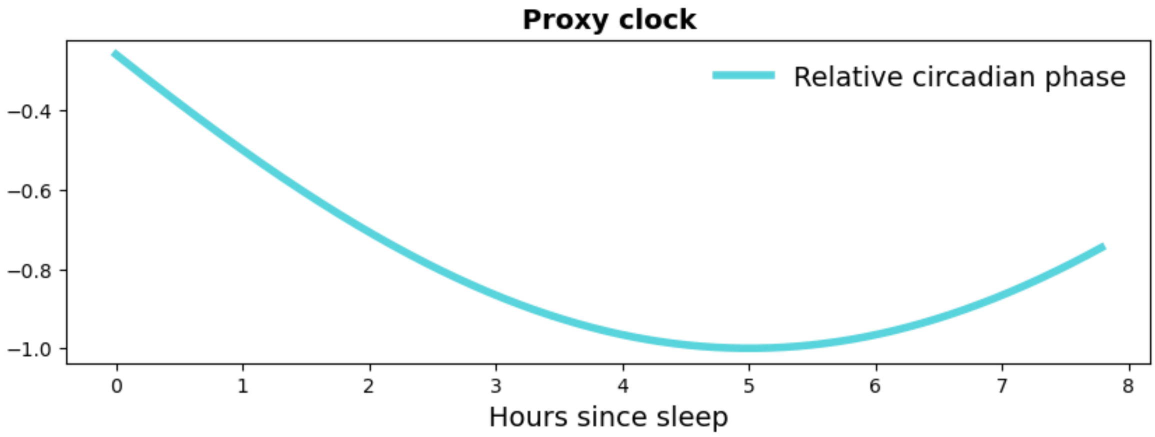

Figure 3: The cosine-shaped proxy clock for modeling the circadian phase during the sleep.

3.4 Labeled sleep

The sleep stages were recorded and determined from polysomnography in 30 seconds intervals. Each stage was given an integer value representing a stage: wake = 0, N1 = 1, N2 = 2, N3 = 3, REM = 5.

The recorded sleep stages are used as the labels in the multi class classification problem. The recorded sleep stages of one test subject are illustrated in Figure 4.

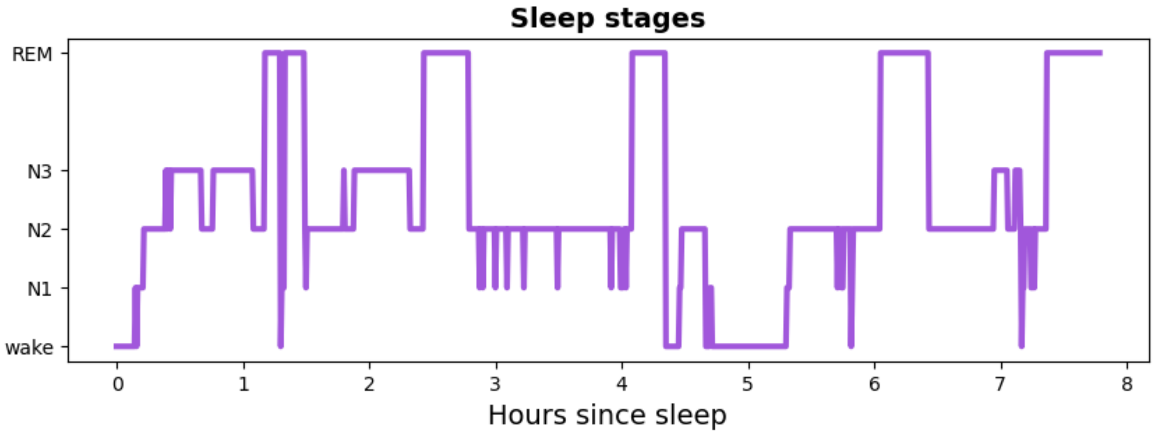

Figure 4: The determined sleep stages recorded from polysomnography. Individual sleep cycle patterns can be identified from the recording.

After proprocessing the data, we were left with 31 subjects and 127405 rows in total. Each row, or data point, represents a 30-second interval containing the subject ID, the features (heart rate, activity, circadian phase), and the labeled sleep stage.

4 Methods

The following sections describes the feature engineering, classification, and evaluation methods used in the study.

4.1 Feature engineering

Predicting the sleep stage proves to be a challenging task, and the three features (HR, activity, circadian phase) alone aren't sufficient for predicting it accurately. To address this, rather than relying on a single row of observations, we attempt to classify the sleep stage at a given time by considering all observations inside a 5 minutes interval beginning 2.5 minutes before and ending 2.5 minutes after the designated measurement.

$$\begin{bmatrix} x_1^{(i)} & x_2^{(i)} & x_3^{(i)} \end{bmatrix} \xrightarrow{\text{extend context}} \begin{bmatrix} x_1^{(i-5)} & x_2^{(i-5)} & x_3^{(i-5)} \\ \vdots & \vdots & \vdots \\ x_1^{(i)} & x_2^{(i)} & x_3^{(i)} \\ \vdots & \vdots & \vdots \\ x_1^{(i+5)} & x_2^{(i+5)} & x_3^{(i+5)} \end{bmatrix} \xrightarrow{\text{reshape } \mathbb{R}^{11\times 3}\to\mathbb{R}^{33}} \begin{bmatrix} x_1^{(i-5)} & \dots & x_3^{(i+5)} \end{bmatrix}$$

In addition to the extended context, the average heart rate, activity count, and the time into the sleep is added as additional features. In total, the feature space was extended from 3 to 36.

The labels are simply processed into three discrete stages, Wake (0), NREM (1), and NREM (2) representing the classes in the classification problem.

4.2 Dimensionality reduction

To increase the computational efficiency and to combat the high dimensionality, the data is projected to a lower dimensional space. Dimensionality reduction was tested using PCA and SVD with varying number of components. Best results were obtained using SVD (singular value decomposition) with ten components. Thus, the number of features is reduced from 36 to 10.

The SVD is calculated using Scikit-learn library's TruncatedSVD function.

4.3 Test train split

The data is split into test and train set using the conventional 20/80 split. In order to mitigate information leakage between the sets, the split is conducted user-wise before the feature engineering phase and thus ensuring that there are no overlapping data between the training and testing sets.

The split was performed using k-fold cross-validation to attain a generalized solution that minimizes the bias from the test train split.

4.4 Classification

Many studies have shown encouraging outcomes utilizing tree-based algorithms (Boostani et al., 2017; Sharma et al., 2021).

In this study, three different gradient boosting implementations, XGBoost (Chen & Guestrin, 2016), LightGBM (Ke et al., 2017), and CatBoost (Dorogush et al., 2018), were employed. Out of the three, CatBoost was able to achieve the most promising results in terms of accuracy across the classes.

CatBoost is an open-source gradient boosting framework that employs a variant of the depth-first tree construction strategy and incorporates regularization techniques to prevent overfitting. Moreover, it has a built-in feature for handling imbalanced datasets by incorporating weights during training. During the training, each class was assigned a weight corresponding to the class' squared inverse frequency.

Thus, for the classification, the CatBoost algorithm was used to perform multi-class classification. The multiclass classifier was fitted with the following parameters:

- depth = 6
- learning_rate = 0.05
- iterations = 100
- auto_class_weights = "SqrtBalanced",

5 Results

The transformed training data was used to fit the gradient boosting classifier, which was then assessed using the test dataset obtained through k-fold cross-validation. The evaluation of classification performance utilized class-wise metrics calculated using scikit-learn. The model's precision, recall, F1 score, and support for all classes as recorded in Table 2.

5.1 Model accuracy

The model demonstrated promising accuracy in distinguishing between wake, NREM, and REM sleep stages, with detailed class accuracies and scores provided in Table 2. The true sleep stages overlayed with the model's predictions are illustrated in Figure 5.

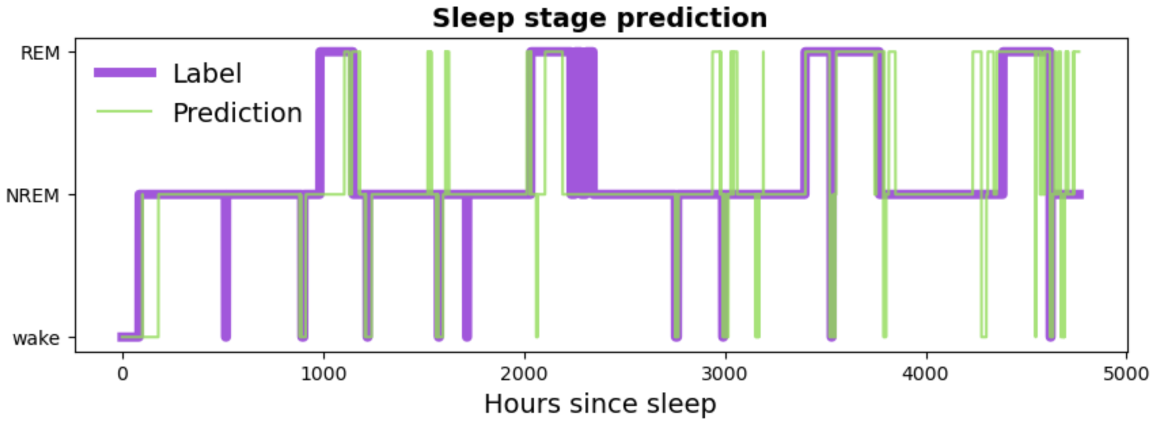

Figure 5: The sleep stage measurements from PSG and the predicted sleep stages using the CatBoost classifier.

Overall, the model demonstrated a consistent ability to predict the NREM phase, while being less accurate in classifying the wake and REM stages. It achieved promising accuracy in classifying NREM periods, attaining a precision of 0.81 and an F1-score of 0.79. However, the accuracy for the Wake and REM classes was relatively lower, at 0.52 and 0.45, accompanied by corresponding F1-scores of 0.55 and 0.47, respectively.

Table 2: Average Metrics Across Five Tables

	Precision	Recall	F1-Score	Support
Wake	0.52	0.60	0.55	2111
NREM	0.81	0.77	0.79	17416
REM	0.45	0.50	0.47	5529
Accuracy			0.70	25057
Macro Avg	0.59	0.62	0.60	25057
Weighted Avg	0.71	0.70	0.70	25057

Despite the model's inability to reliably classify REM phases, it is reassuring to note that the majority of false negatives are attributed to NREM (Figure 6). The classification of REM stages as "Wake" stands at only 3%, which is arguably a less favorable outcome compared to the misclassification of NREM.

However, it is noteworthy that over a fifth of the wake stages were inaccurately classified as REM stage, indicating a notable challenge in accurately distinguishing between these two sleep stages.

The results can be further interpreted using the ROC-curve (Receiver operating characteristic) for each class. The use of micro-averaged ROC curve is motivated by the class imbalance of the multi class classification problem. The Micro-averaged One-vs-Rest curve has the ability to assess the performance across all classes by aggregating true positive, false positive, and false negative rates. The results are shown in Figure 6. With an AUC (area under curve) value of 0.89, the model demonstrates a promising ability to classify the sleep stages.

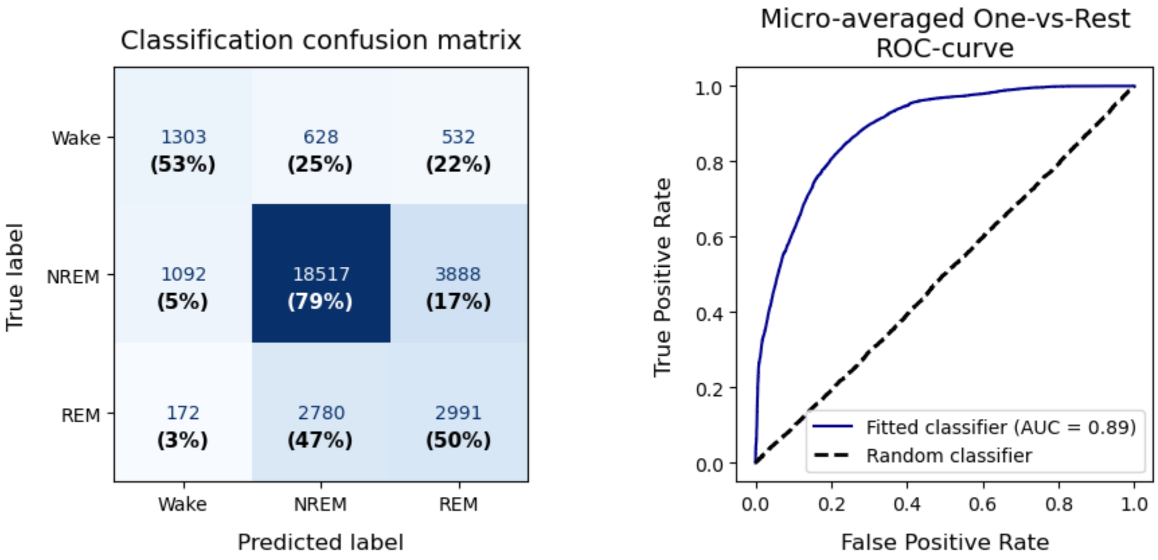

Figure 6: The confusion matrix (left) summarises the amount of correct and incorrect predictions per class. The confusion matrix of a perfect classifier would be a diagonal matrix. The micro-averaged one-vs-rest (right) ROC-curve for assessing the classifier's performance at different threshold levels.

5.2 Feature importance

Some features carry more significance than others. In order to assess the usefulness of each feature, the feature importance value of each feature can be extracted from fitted the classifier.

However, during the dimensionality reduction stage, the original features were decomposed to only ten components. Thus, the feature importance of these components had to be remapped to the original features inf order for us to attain useful information about the importance of the different measurements.

$$X \xrightarrow{\text{Dimensionality reduction}} X_{SVD} \xrightarrow{\text{Train}} \boxed{\text{classifier}} \xrightarrow{\text{Feature importance}} F_{X,svd} \xrightarrow{\text{Map to original features}} F_X$$

The transformed features importance values were mapped into the original values by identifying the top contributing principal components and aggregating their the loadings. The expected feature importance of the original 36 features are presented in Figure 7.

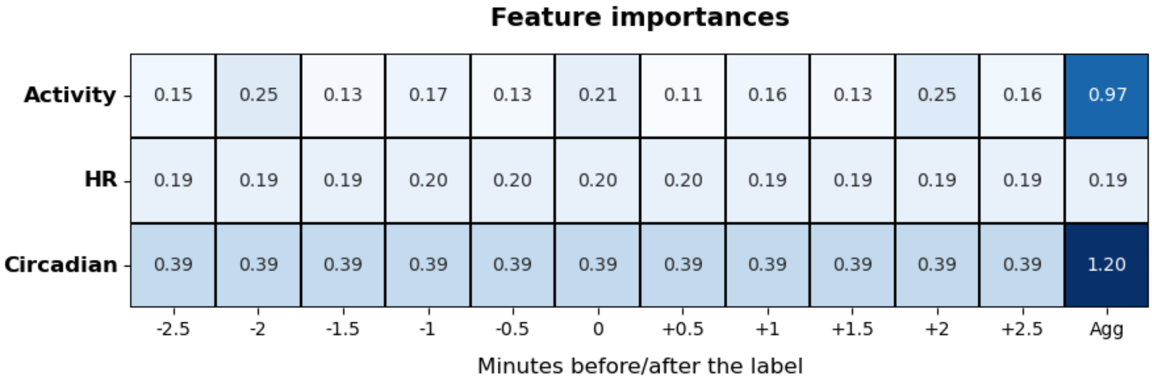

Figure 7: The feature importance of the original features. The aggregated measures, in particular the total activity during the time period and the relative time into sleep (bottom right), appeared to be the most important features.

Based on the figure, the most important features seemed to be the total number of activities during the 5-minute time period (0.97) and the relative time of the measurement (1.2).

Surprisingly, the significance of activity and heart rate variability as features was not particularly high, whereas the circadian phase proved to provide valuable information. While the circadian phase was important, the score didn't appear to be influenced by the relative time of the measurement. It might signify that the sleep stage is closely related to the absolute phase of the circadian rhythm.

Additionally, measurements taken after the label showed slightly greater importance compared to those taken before the label. As one might intuitively expect, classifying the sleep stage during the stage itself proved to be easier than doing so before it.

6 Conclusion & Discussion

The aim of this study was to assess whether reliable sleep stage classification could be achieved using a consumer wearable device, specifically the Apple Watch, in a non-clinical setting. The study explored the feasibility of using a gradient boosting classifier CatBoost for predicting the labeled sleep stages.

The results demonstrated that the model can predict NREM sleep with reasonable reliability with an overall accuracy close to 70%. However, it struggles with REM and Wake stages, only achieving an accuracy of around 50%.

Misclassifications between Wake and REM stages pose a significant challenge, suggesting that additional features crafted through feature engineering or new types of measurements are needed for improving the accuracy.

The project highlighted the potential of consumer wearables in sleep stage classification, offering a less intrusive and more accessible alternative to PSG. The activity logs and the circadian phase were found to be useful in predicting sleep stages. However, the lack of importance on heart rate variability was an unexpected outcome.

6.1 Limitations and further studies

Despite the promising results, the study had limitations. The labeled data was fully dependent on the labeled PSG values which might contain errors. Moreover, the sample of subjects at the University of Michigan might also introduce potential biases.

Another limitation was the small sample of 31 persons, limiting the number of ways to split the dataset into training ad testing set. By splitting the dataset by the users, the class balances fluctuated a bit each fold of the cross validation. Furthermore, some recordings were interpolated during the data preprocessing step, which might also skew the results if not handled properly.

Future research could focus on improving classifier precision for REM sleep, using new types of measurements, such as breathing intervals, and integrating additional context-aware features such as the stress level of the subject.

Deep learning techniques, such as recurrent neural networks (RNN), could be employed to capture more complex patterns and refine predictions. Moreover, gathering real-world sleep data will be valuable in advancing the models' generalizability.

7 Source code

The source code can be found from the project's GitHub repository: `https://github.com/jiemingyou/sleep-stage-prediction-bst/tree/main`

References

Boostani, R., Karimzadeh, F., & Nami, M. (2017). A comparative review on sleep stage classification methods in patients and healthy individuals. *Computer Methods and Programs in Biomedicine, 140*, 77-91. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0169260716308276` doi: https://doi.org/10.1016/j.cmpb.2016.12.004

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining.* ACM. Retrieved from `http://dx.doi.org/10.1145/2939672.2939785` doi: 10.1145/2939672.2939785

Dorogush, A. V., Ershov, V., & Gulin, A. (2018). *Catboost: gradient boosting with categorical features support.* arXiv. Retrieved from `https://arxiv.org/abs/1810.11363` doi: 10.48550/ARXIV.1810.11363

Forger, D. B., Jewett, M. E., & Kronauer, R. E. (1999, December). A simpler model of the human circadian pacemaker. *Journal of Biological Rhythms, 14*(6), 533–538. Retrieved from `http://dx.doi.org/10.1177/074873099129000867` doi: 10.1177/074873099129000867

Gaiduk, M., Serrano Alarcón, , Seepold, R., & Martínez Madrid, N. (2023, July). Current status and prospects of automatic sleep stages scoring: Review. *Biomedical Engineering Letters, 13*(3), 247–272. Retrieved from `http://dx.doi.org/10.1007/s13534-023-00299-3` doi: 10.1007/s13534-023-00299-3

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems, 30*, 3146–3154.

Miettinen, T., Myllymaa, K., Westeren-Punnonen, S., Ahlberg, J., Hukkanen, T., Toyras, J., ... Myllymaa, S. (2018, July). Success rate and technical quality of home polysomnography with self-applicable electrode set in subjects with possible sleep bruxism. *IEEE Journal of Biomedical and Health Informatics, 22*(4), 1124–1132. Retrieved from `http://dx.doi.org/10.1109/JBHI.2017.2741522` doi: 10.1109/jbhi.2017.2741522

Patel, A. K. (2022, sep 7). *Physiology, Sleep Stages.* https://www.ncbi.nlm.nih.gov/books/NBK526132/.

Sharma, M., Tiwari, J., & Acharya, U. R. (2021, March). Automatic sleep-stage scoring in healthy and sleep disorder patients using optimal wavelet filter bank technique with eeg signals. *International Journal of Environmental Research and Public Health, 18*(6), 3087. Retrieved from `http://dx.doi.org/10.3390/ijerph18063087` doi: 10.3390/ijerph18063087

Sleep Phases and Stages. (2022, mar 24). https://www.nhlbi.nih.gov/health/sleep/stages-of-sleep. ([Online; accessed 2023-11-16])

Walch, O. (2020). *Motion and heart rate from a wrist-worn wearable and labeled sleep from polysomnography.* PhysioNet. Retrieved from `https://physionet.org/content/sleep-accel/` doi: 10.13026/5OYH-TT09