## Introduction

In the current fast-paced world, people tend to overlook the necessity of getting proper sleep. Insomnia, sleep-disordered breathing, and other sleep disorders are common both for adults and children [1], so the importance of measuring our sleep is increasing, as it can foster the recognition and treatment of such conditions, thus improving our overall sleep quality. In the past it was possible only using Polysomnography (PSG), which is an advanced tool, but is obtrusive and available only in a sleep lab with a sleep technician, and thus is not suitable for sleep measurements taking longer than one or two nights [2]. With the popularity and accessibility of wearable devices such as smartwatches or smartbands, there have been a lot of new possibilities introduced in the domain of sleep measurement. One of the parameters that these devices can measure is the length of our sleep, as well as its different stages which can be divided into: rapid eye movement (REM) and non-rapid eye movement (NREM) phase, which can be further divided into three stages, N1-N3 [3]. NREM stages constitute about 75% of total sleep, with most time spent in the N2 stage [3].

Olivia Walch et al. [2] created an application to collect raw acceleration data and heart rate from an Apple Watch, which they used to develop classifiers that were then compared to the polysomnography data. Their research aims to validate the effectiveness of consumer wearables, like the Apple Watch, in accurately monitoring sleep patterns, particularly differentiating between wake, NREM, and REM sleep stages. The data collected by them is available for everyone and this project aims to analyze the data to develop machine learning models for sleep stages classification and to identify conditions that promote longer, more restorative deep sleep phases.

Several machine learning algorithms were used for this purpose, including logistic regression, random forest and neural networks, with the last one performing the best. Classification between awake and asleep states was investigated, as well as classification between wake, sleep and deep sleep phases. The first classifiers performed well, and their performances were close to Walch's et al. [2], while the second was struggling with making accurate predictions.

Data analysis showed weak and moderate correlations between activity routine like the number of steps taken every day and different parameters of a person's sleep, obtained from the Apple Watch data [4].

## Problem Formulation

Sleep quality is a complex concept, that can be defined on both subjective and objective levels [5]. While tools like the Pittsburgh Sleep Quality Index (PSQI) [6] offer subjective assessments through self-reporting questionnaires, objective measures such as the distribution of sleep phases provide a more quantifiable approach to understanding sleep quality. The deep sleep phase, or NREM stage 3, is particularly crucial for overall health and well-being, with insufficient deep sleep linked to various disorders, including dementia [5]. It is referred to as the most restorative of the sleep stages [7]. A lot of other aspects can be considered important when discussing the meaning of sleep quality, but all of the aforementioned emerge as relevant to the dataset in consideration.

The focus of this project is to analyze sleep data from a study by Walch et al. [2] and develop machine learning models capable of accurately predicting sleep states using Apple Watch data, specifically focusing on two aspects:

1. **Sleep/Wake Classification**: Developing a model to distinguish between sleep and wake states, using heart rate and acceleration data from a smartwatch.
2. **Deep Sleep Phase Identification**: Creating a model to identify the deep sleep phase (N3 stage). This aspect is important, as deep sleep is essential for health, and such model can provide valuable insights into sleep quality.

Additionally, the study seeks to explore correlations between daytime physical activities, such as step count, and sleep quality. This could offer new insights into how lifestyle choices impact sleep health.

## Dataset Description

The dataset contains acceleration and heart rate data collected from Apple Watch, along with labeled sleep recorded from gold-standard polysomnography. The data were collected from 31 subjects. They are stored in multiple .txt files, each file corresponding to a different subject and a different measurement. A detailed description of the data is presented below:

- **Motion (Acceleration)**: captured by the Apple Watch's sensors, includes measurements of acceleration along three axes: x, y, and z. Each axis's acceleration is expressed in 'g', a unit of gravitational force. Each datapoint provides a timestamp in seconds since the start of the PSG study, followed by the x, y, and z acceleration values.
- **Heart Rate (BPM)**: also recorded by the Apple Watch. Each datapoint contains a timestamp (in seconds from the start of the PSG) and the corresponding heart rate in beats per minute (bpm).
- **Steps (Count)**: represents the count of steps taken by the subject, as recorded by the Apple Watch. Each datapoint includes a timestamp (seconds since the PSG start) and the total number of steps counted in the time bin from this timestamp to the next.
- **Labeled Sleep (Sleep Stages)**: data recorded through polysomnography. Each datapoint provides a timestamp (in seconds since the start of PSG) and a sleep stage label. The sleep stages are numerically coded as 0-5, where 'wake' is represented by 0, N1 by 1, N2 by 2, N3 by 3, and REM sleep by 5.

***Table 1.*** *Description of variables in the dataset.*

| Variable | Type | Description |
|---|---|---|
| Motion (Acceleration) | Quantitative | Measurements of acceleration along x, y, z axes in 'g', recorded from an Apple Watch. Timestamped in seconds since PSG start. |
| Heart Rate (BPM) | Quantitative | Heart rate data in beats per minute, recorded from an Apple Watch. Timestamped in seconds since PSG start. |
| Steps (Count) | Quantitative | Count of steps taken, as recorded by an Apple Watch. Timestamped in seconds since PSG start, with total steps counted in each time bin. |
| Labeled Sleep | Categorical | Sleep stage data from polysomnography, labeled as stages 0-5 (wake to REM). Timestamped in seconds since PSG start. |

First step was preprocessing the data and making it usable and interpretable. As all data, except for polysomnography recordings, were recorded with ambiguous time stamps, it was necessary to unify the data to have corresponding time stamps. Starting with the data that are relevant during PSG, that is heart rate and motion, they were segmented into 30-second bins. For both, the averages over 30 seconds periods were used to form new data points. For motion data, additional variable of magnitude was added, which corresponds to the length of the vector of acceleration. To preserve information about the variability of the data, standard deviation of heart rate and motion data for each time stamp was added.

The steps count data were treated differently. As step count is not important during the night, but rather the number of steps taken during the whole preceding day can matter, small-sized bins were not necessary. Hourly bins were used instead, to simplify working with the data, at the same time keeping the important information like distribution of steps count over the day.

The processed data turned out to have 11 data points with heart rate standard deviation missing. These data points were discarded. There were also data points with incorrect labels like -1 or 4, all of which were discarded. The resulting dataset had 25141 data points.

Next, counts for the classes to be classified were checked. The numbers of data points with each label of the sleep stage are shown in *Table 2* and with the labels of sleep/wake are presented in *Table 3*.

***Table 2.*** *Sleep stage labels counts.*

| Label | Count |
|---|---|
| 0 | 2166 |
| 1 | 1761 |
| 2 | 12483 |
| 3 | 3189 |
| 5 | 5542 |

*Table 3. Sleep/wake labels counts.*

| Label | Count |
|---|---|
| 0 (sleep) | 22975 |
| 1 (wake) | 2166 |

The most data points have label 2, which reflects the fact, that we spend the significant amount of time (around 45%) in the N2 sleep stage [3]. As one can observe, there is a big difference in numbers of instances between the wake and the sleep class, which may cause problems in developing the classification models. This issue will be addressed later, in the *Methods* section.

After completing the initial preprocessing, more interpretable data for analysis needed to be produced. In the next steps, the impact of daytime physical activity on various aspects of sleep quality was to be investigated. To make data more interpretable, several actions were taken.

The sum of steps count in a chosen time range for every subject was calculated. Various time ranges could be considered for counting the steps taken, but initially the data were

aggregated from 2 hours before the start of the PSG recording. The reason behind this choice is that research shows that walking before bedtime can have a positive effect on sleep quality [9].

Sleep duration was obtained by counting all the time stamps with the sleep stage labeled as 1, 2, 3 or 5, corresponding to N1, N2, N3 and REM phases. Then the resulting number was multiplied by 30, to obtain the duration of sleep in seconds, as the data from PSG was recorded every 30 seconds. After that it was divided by 3600, as that is the number of seconds in an hour, to obtain the sleep duration in hours.

Deep sleep duration was also examined, as it can be associated with better sleep quality and has a significant role in restoration of our bodies. The data was produced in a similar manner to sleep duration data. This time, the data points with the sleep phase classified as 3 were counted, as it corresponds to N3 sleep stage.

Additionally, the ratio of the deep sleep duration to the overall sleep duration was considered. It was obtained by dividing deep sleep duration by sleep duration for each subject and then multiplied by 100, to convert it into percentage values, for easier analysis.

Finally, the number of arousals during the night for each subject was calculated. It was done by searching for the changes from asleep to awake state, and then checking the length of the identified nocturnal awakening. According to a study by Michael Winser et al. [8], the length of an awakening to be remembered by a person must be at least on average 4 minutes and 19 seconds, thus arousals over 4 minutes long were counted.

# Methods

## Data Analysis

After obtaining more interpretable data, the distributions were examined. For that, histograms were used. First, the distribution of steps count was plotted.

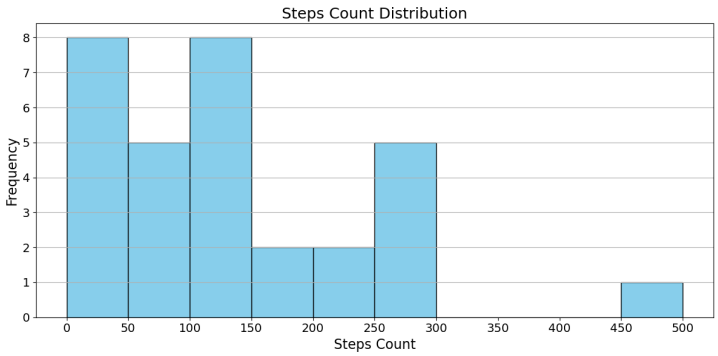


***Figure 1.*** *Steps count distribution.*

One can observe that people did not tend to take a lot of steps close to bedtime. A possible outlier can be spotted, as one data point seems quite distant from the majority.

Next, the sleep duration distribution was plotted.

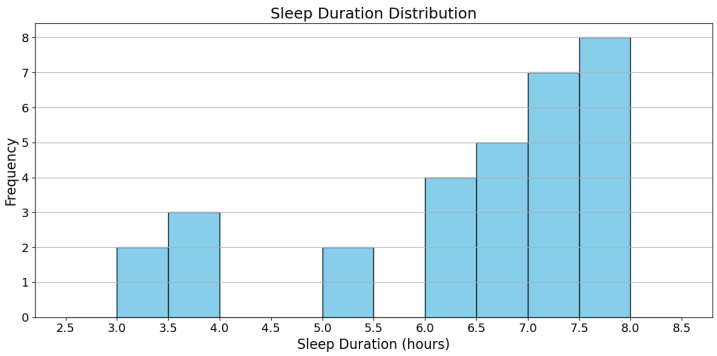


***Figure 2.*** *Sleep duration distribution.*

According to the National Sleep Foundation's sleep time duration recommendations, an adult should get 7 to 9 hours of sleep [10]. As one can see from the distribution, around half of the subjects managed to get the appropriate sleep. One can also observe that there are some subjects who got less than 4 hours of sleep during the PSG. After investigating the original data carefully, it turns out that for these subjects, PSG data was recorded for shorter time. In the dataset description [4] it is mentioned that some of the Apple Watches ran out of battery and the data were cropped. This is most likely the case for the outlying subjects, so they were not included in the further analysis (however, their data was still later used for classification models).

The distribution of deep sleep duration was then plotted.

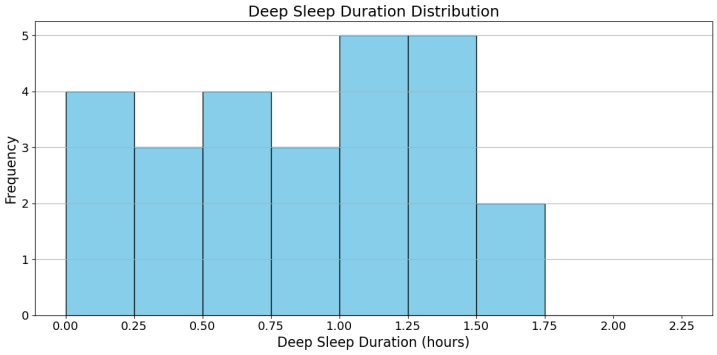


***Figure 3.*** *Deep sleep duration distribution.*

There are no official recommendations for the amount a person should spend in the deepest phase of the sleep. However, it should take around 25% of the sleep [3] and considering the official norms for the sleep duration of an adult person, which is 7 to 9 hours [10], such sleep should take around 105 minutes (1.75 hours) to 135 minutes (2.25 hours). The distribution clearly shows that most of the subjects did not fulfill this recommendation.

Next, the distribution of deep sleep percentage was examined.

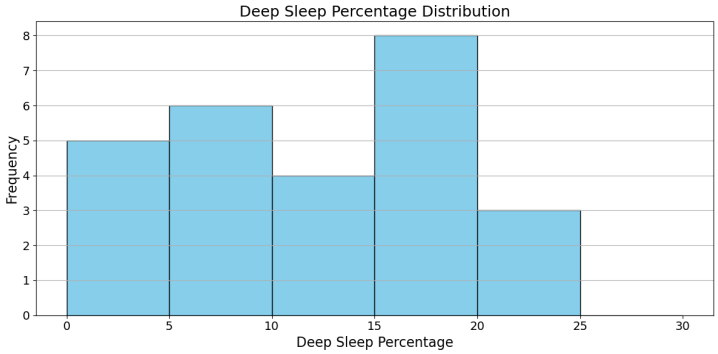


***Figure 4.*** *Deep sleep percentage distribution.*

The distribution shows that, while there are subjects that spend the right amount in the deepest sleep phase, there is a significant number of people who had far too little ratio of deep sleep to the sleep duration. However, it might result from some other nuances of collecting the data that were not discovered during the preceding analysis.

Finally, the distribution of nocturnal awakenings count was investigated.

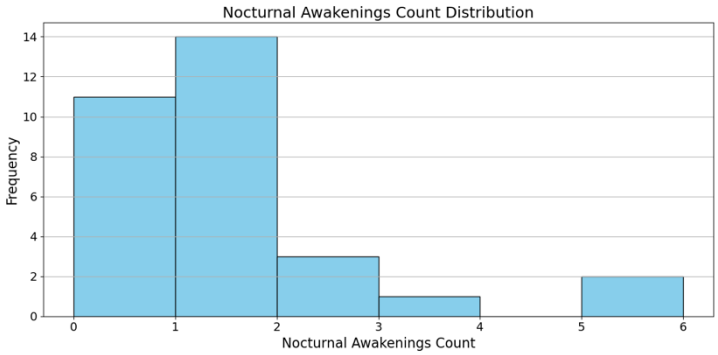


***Figure 5.*** *Nocturnal Awakenings Count Distribution.*

The distribution shows that for most of the subject at least one nocturnal awakening occurred during the night. Research considers such number a norm [8], so none of the subjects showed any anomaly.

Next, relations of the sleep characteristics with the steps counts were plotted, to search for correlations. For that scatter plots were used, along with appropriate correlations coefficients. To determine the correct coefficients to use, the Shapiro-Wilk test was conducted to check the normality of the data. Then, based on the result of the test, the proper coefficients were chosen. The results of the Shapiro-Wilk test are presented in the *Table 4*:

***Table 4.*** *The results of Shapiro-Wilk tests.*

| Metric | Shapiro-Wilk Test Statistic | p-value |
|---|---|---|
| Steps Counts | 0.884570 | 0.007199 |
| Sleep Duration | 0.917419 | 0.039131 |
| Deep Sleep Duration | 0.951142 | 0.246587 |
| Deep Sleep Percentage | 0.962960 | 0.453132 |
| Awakenings Count | 0.741148 | 0.000020 |

Using a threshold of $p > 0.05$ it was inferred that deep sleep duration and deep sleep percentage can be treated as distributed normally. The steps counts, sleep duration data and nocturnal awakenings count do not have normal distributions. The choice for the correlation coefficients was between Pearson's correlation and Spearman's correlation, as they are the main statistical tools for assessing the strength and direction of a linear relationship between two continuous variables. Pearson's correlation coefficient is most effective for data sets that are normally distributed and linearly related, while Spearman's correlation is more appropriate for data that may not be normally distributed or is ordinal in nature. This distinction is critical because using the wrong type of correlation coefficient could lead to inaccurate conclusions about the relationship between sleep characteristics and step counts.

Given the results of the Shapiro-Wilk test and the fact that steps count relation with other variables is considered, Spearman's correlation was selected for all analyzing all of the variables, due to the non-normal distribution of steps count, as evidenced by a p-value less than 0.05.

After choosing the correct correlation coefficients, the scatter plots were produced.

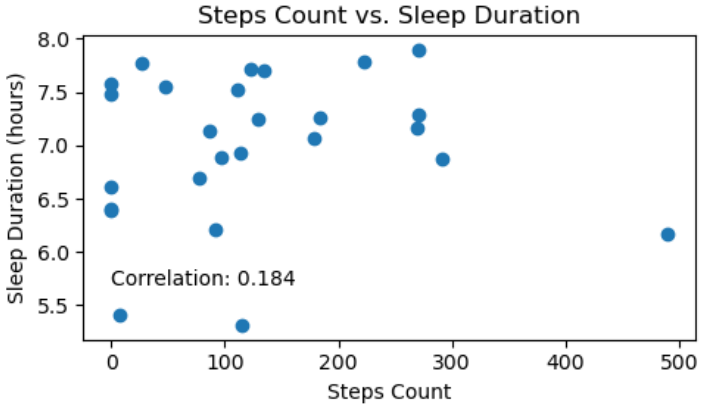


***Figure 6.*** *Steps Count vs. Sleep Duration.*

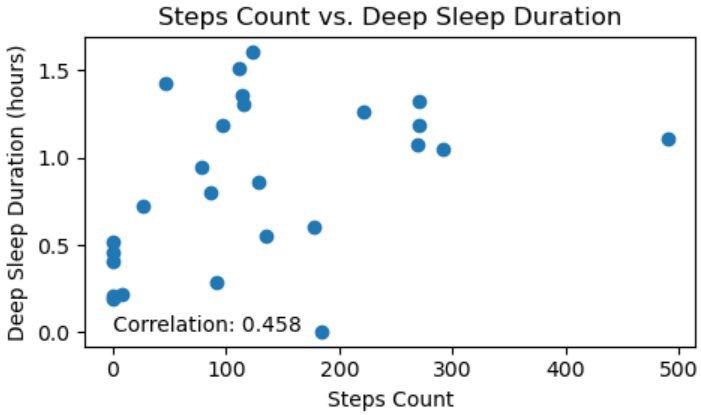


***Figure 7.*** *Steps Count vs. Deep Sleep Duration.*

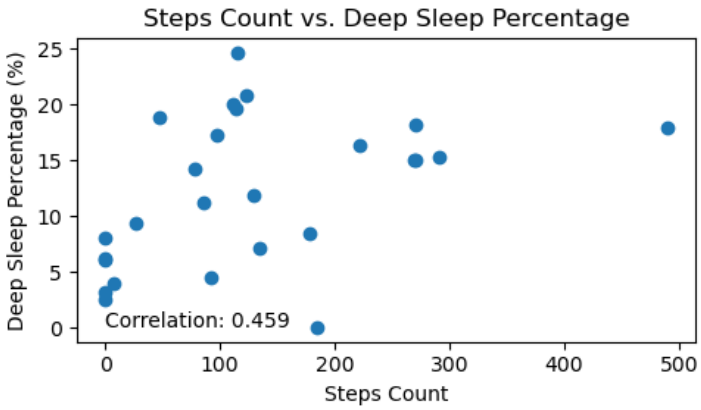


***Figure 8.*** *Steps Count vs. Deep Sleep Percentage.*

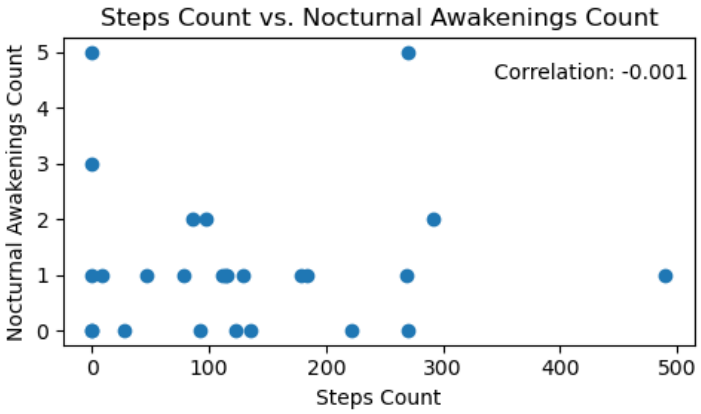


***Figure 9.*** *Steps Count vs. Nocturnal Awakenings Count.*

As one can observe, there are some correlations between the variables, which might suggest the impact of taking more steps before bed on the sleep quality. There are moderate correlations between steps count and deep sleep duration and percentage. This topic will be further brought up in *Conclusions & Discussion*.

The same steps were repeated but aggregating the steps from the preceding day (counting from 18 hours before PSG till PSG), and from the whole period (counting from 74 hours before PSG till PSG, some subjects had data recorded starting from even over 100 hours before PSG, but 74 hours was the minimum for all subjects and was chosen not to distort the distribution). The correlations were a lot weaker, however, some interesting observations could be made. *Figure 10* presents the relation between steps count and nocturnal awakenings count with steps aggregated from the day preceding the PSG.

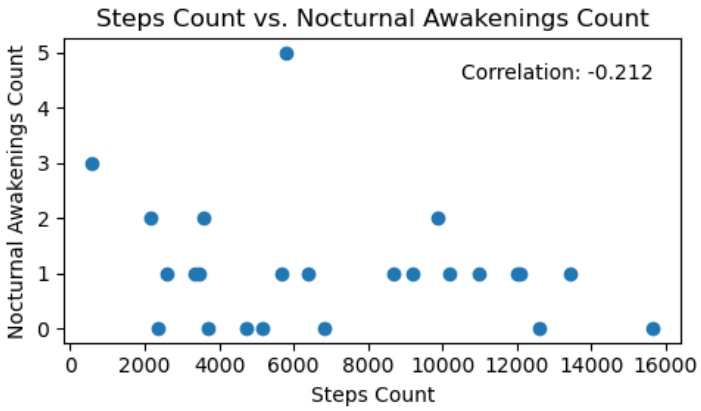


***Figure 10.*** *Steps Count vs. Arousals Count with steps aggregated from the day preceding PSG.*

There is a weak negative correlation between these two variables, which suggests that there might be a connection between the number of times a person wakes up at night and the number of steps they take. The people who took more steps generally have lower arousals count.

To further explore the correlations, the subjects with the biggest and the smallest numbers of steps were examined. Their number of steps made every hour were plotted on lineplots.

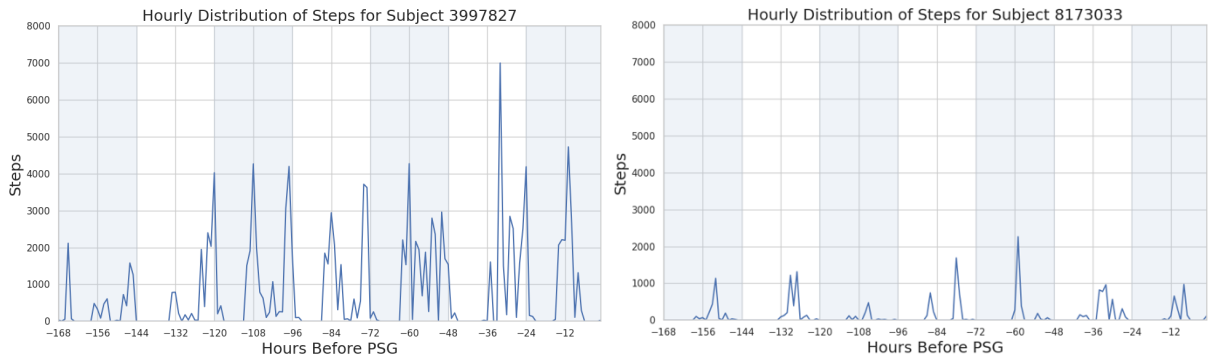


***Figure 11.*** *Hourly Distribution of Steps for Subjects 3997827 and 8173033*

These plots show a major difference in subjects' activity routines. Subject 3997827 does a lot of steps daily, while 8173033 does much less. *Table 5* presents sleep statistics for both subjects:

***Table 5.*** *Comparison of Sleep Statistics for Subjects 3997827 and 8173033*

| **Subject** | 3997827 | 8173033 |
|---|---|---|
| **Average steps per day** | 18614.14 | 2838.29 |
| **Hours of sleep** | 7.775 | 7.72 |
| **Hours of deep sleep** | 0.725 | 1.61 |
| **Percentage of deep sleep** | 9.32 | 20.84 |
| **Nocturnal awakenings** | 0 | 0 |

These data clearly show that the number of steps does not have to directly affect our sleep quality. Subject 8173033 did not a lot of steps daily and had a bigger percentage of deep sleep, with no nocturnal awakenings, and thus their sleep is considered better.

## Machine Learning

After exploring the data, the models for sleep/wake classification were developed. As the features, heart rate and motion data with their standard deviations were used. The chosen classifiers included logistic regression, random forest, and neural networks, each selected for their unique strengths and capabilities in handling complex datasets.

1. **Logistic regression** was chosen for its efficiency and interpretability.
2. **Random forest** was employed, because of its suitability for multi-class classification problems. Random Forest can capture complex relationships in the data, making it a good candidate for this task.
3. **Neural networks** have the ability to learn non-linear relationships and intricate patterns in data, they are capable of capturing complex interactions between heart rate and motion data that simpler models might miss.

For implementing the chosen classifiers, scikit-learn library for python was used. Each classifier was evaluated using Stratified K-Fold Cross-Validation with five folds, ensuring that each fold was a good representative of the whole. For each fold, the training data was oversampled using SMOTE and then standardized the features using scikit-learn's StandardScaler.

The data was first split into training and testing subsets, by performing a random split. For that scikit-learn's function 'train_test_split' was used, with the stratification based on the target variable. The ratio of the sets was 80% for training and 20% for testing. Then, to properly evaluate the model's performance, cross-validation approach was employed, specifically, StratifiedKFold cross-validation, which preserved the percentage of samples for each class in each fold. The training set was split 5 times for each classifier into the actual training and validation sets.

Due to large classes differences, where wake was represented by n=1807 data points and sleep by n=19665 data points, Synthetic Minority Over-sampling Technique (SMOTE) was used to improve the balance between the classes. It provided a more balanced dataset, resulting in classification metrics that are more trustworthy and comprehensible.

Training the models involved tuning hyperparameters. The goal was to balance the complexity of the model allowing for capturing the patterns in the data, while keeping the reasonable amount of generalization. After experimenting, hyperparameters that ensured the model does not overfit were chosen.

For Logistic Regression, the maximum of 1000 iterations for convergence was selected. Random Forest classifier involved adjusting parameters such as the number of trees in the forest and the maximum depth of the trees. The model was configured with 100 estimators and a maximum depth of 6. For the neural network hyperparameters like number of layers and neurons for each layer, strength of the L2 regularization term, and maximum number of iterations were adjusted. The MLP neural network was structured with two hidden layers containing 50 and 25 nodes, respectively, and employed early stopping with a patience of 20

iterations and an alpha value of 0.001 for regularization. Maximum number of iterations was set to 1000.

After training each model, it was made sure, that the model is not overfitting, by comparing the training sets and testing sets error. A noticeable difference between those metrics is usually a sign of overfitting, so they were kept track of.

Additionally, classifying between wake, sleep, and deep sleep was investigated, however only neural networks were considered as the other two algorithms performed insufficiently. The classifier was trained with the same hyperparameters as for the sleep/wake classification and using the same methods like SMOTE and Stratified K-Fold Cross-Validation.

## Results

Comparison of the trained machine learning models is presented in *Table 6*.

***Table 6**. Comparison of Machine Learning Models.*

| Model | Average CV Accuracy | Sleep Precision | Sleep Recall | Sleep F1-Score | Wake Precision | Wake Recall | Wake F1-Score |
|---|---|---|---|---|---|---|---|
| Logistic Regression | 89.04% | 0.94 | 0.94 | 0.94 | 0.4 | 0.4 | 0.4 |
| Random Forest | 88.44% | 0.96 | 0.92 | 0.94 | 0.4 | 0.54 | 0.46 |
| Neural Network | 89.76% | 0.97 | 0.91 | 0.94 | 0.42 | 0.72 | 0.53 |

Neural network outperforms other models in almost all of the metrics, however, the differences are not that noticeable. All models have high precision, recall, and F1-scores for the sleep state, meaning they are generally good at predicting sleep states correctly. The relatively high recall for wake in the neural network model (0.71) indicates that it is relatively better at identifying actual wake instances compared to the other models, however low value for wake precision (0.42) shows that not a lot of data points classified as wake were actually wake.

Given these results, the neural network seems to be the best model among the three, especially for identifying wake states, although it does tend to predict more wake states than actually exist (as indicated by the lower precision but higher recall).

Neural network used for classifying between wake, sleep and deep sleep scored 78.50% accuracy. Other metrics are presented in *Table 7*.

***Table 7.** Neural Network Performance in Wake/Sleep/Deep Sleep Classification*

| Class | Precision | Recall | F1-Score |
|---|---|---|---|
| 0 (Wake) | 0.43 | 0.66 | 0.52 |
| 1 (Sleep) | 0.93 | 0.82 | 0.88 |
| 3 (Deep Sleep) | 0.60 | 0.83 | 0.7 |

Precision for the wake class is again low, however, recall equals to 66%, so the model is reasonably good at detecting wake instances. For the deep sleep class, precision is under 60%, suggesting that a bit less than half of the deep sleep predictions made by the model are incorrect. The model is, however, quite effective at identifying deep sleep instances, as it correctly predicts 83% of them.

## Conclusion & Discussion

The analysis of the data showed correlations between the activity of a subject, measured in step counts, and their sleep quality. Correlations between steps count shortly before bedtime with deep sleep duration and deep sleep percentage were moderate (around 0.46), suggesting that the activity before bedtime can help a person to maintain a higher quality of their sleep. Analyzing the activity over a longer period did not show any noticeable patterns except for a weak negative correlation between the steps taken during the day preceding the PSG and number of nocturnal awakenings. This might suggest that daytime activity might influence sleep quality. However, there are subjects that contradict this observation, so more thorough analysis, possibly on a bigger group should be conducted to yield more clear results.

The machine learning models developed for classifying sleep and wake state are satisfactory and provide a good tool for predicting a person's sleep state, especially neural network. However, the models still suffer from low scores for wake class, and collecting or generating more heart rate and motion data might allow for improvements in this.

The neural network's performance in identifying the deep sleep phase, though not as high as in sleep/wake classification, is still notable. The model showed a reasonable ability to detect deep sleep phases, which is crucial for assessing sleep quality and identifying potential sleep disorders.

The main limitation of this work was the significant class imbalance in the dataset, especially in the wake/sleep classification, it affected the model's performance and generalizability. Also, the number of subjects in the dataset is not big and their background is unknown, which could distort the true correlations because of the lack of diversity and representativeness in the sample. This is particularly relevant in sleep studies, where factors like age, gender, lifestyle, and health conditions can significantly influence sleep patterns [3].

The resulting models could be used for further analysis of the sleep quality, producing predictions for other data, however, the dataset does not provide sufficient amount of data to perform such analysis. Future work could focus on collecting more data from Apple Watch, which could be used for classifying if a person is asleep or in a deep sleep phase, which then could be analyzed. Incorporating data from a wider range of devices and demographic groups could improve the models' robustness and applicability too. Future analysis might also include conducting longitudinal studies to assess the impact of lifestyle changes on sleep patterns over time. Exploring integration of these models with health platforms could provide improvements in comprehensive health monitoring and feedback.

The effectiveness of Apple Watches in monitoring sleep stages highlights the potential of consumer wearable devices in personal health management and sleep research. The potential of machine learning and wearable devices in enhancing our understanding of sleep patterns and quality is evident, paving the way for more personalized and accessible sleep health management.

## References


1. Karna, B., Sankari, A., Tatikonda, G. (2023) Sleep Disorder. Retrieved from: https://www.ncbi.nlm.nih.gov/books/NBK560720/
2. Walch, O., Huang, Y., Forger, D. B., & Goldstein, C. (2019). Sleep stage prediction with raw acceleration and photoplethysmography heart rate data derived from a consumer wearable device. *Sleep*, Volume 42, Issue 12. https://doi.org/10.1093/sleep/zsz180
3. Patel, A. K., Reddy, V., Shumway, K. R. et al. (2022). Physiology, Sleep Stages. Retrieved from: https://www.ncbi.nlm.nih.gov/books/NBK526132/
4. Motion and heart rate from a wrist-worn wearable and labeled sleep from polysomnography https://physionet.org/content/sleep-accel/1.0.0/labels/#files-panel
5. Nelson, K. L., Davis, J. E., & Corbett, C. F. (2021). Sleep quality: An evolutionary concept analysis. *Nursing Forum*, *57*(1), 144–151. https://doi.org/10.1111/nuf.12659
6. Buysse, D. J., Reynolds, C. F., Monk, T. H., Berman, S. R., & Kupfer, D. J. (1989). The Pittsburgh sleep quality index: A new instrument for psychiatric practice and research. *Psychiatry Research*, *28*(2), 193–213. https://doi.org/10.1016/0165-1781(89)90047-4
7. Himali, J. J. et al. (2023). Association between Slow-Wave sleep loss and incident dementia. *JAMA Neurology*, *80*(12), 1326. https://doi.org/10.1001/jamaneurol.2023.3889
8. Schade, M. M. et al. (2020). Enhancing Slow Oscillations and Increasing N3 Sleep Proportion with Supervised, Non-Phase-Locked Pink Noise and Other Non-Standard Auditory Stimulation During NREM Sleep. *Nature and Science of Sleep*, *Volume 12*, 411–429. https://doi.org/10.2147/nss.s243204
9. Winser, M. A., McBean, A. L., & Montgomery-Downs, H. E. (2013). Minimum duration of actigraphy-defined nocturnal awakenings necessary for morning recall. *Sleep Medicine*, *14*(7), 688–691. https://doi.org/10.1016/j.sleep.2013.03.018
10. Bisson, A. N., Robinson, S. A., & Lachman, M. E. (2019). Walk to a better night of sleep: testing the relationship between physical activity and sleep. *Sleep Health*, *5*(5), 487–494. https://doi.org/10.1016/j.sleh.2019.06.003
11. Hirshkowitz, M. et al. (2015). National Sleep Foundation's sleep time duration recommendations: methodology and results summary. *Sleep Health*, *1*(1), 40–43. https://doi.org/10.1016/j.sleh.2014.12.010