Classification of Sleep and Wake States from Apple Watch and Polysomnography Data with Logistic Regression and Random Forests

Introduction

On average people spend a third of their life sleeping. It is a major part of our lives and because of that it is important to understand how we sleep. There are many factors to consider, including sleep time, sleep quality and sleep architecture. Sleep monitoring has emerged as a major aspect of health assessment, which gives clues about the well-being of a person. As technology advances, the integration of machine learning techniques in sleep stage classification will be more apparent. This report will aid with the development of machine learning algorithms by using data from the Apple Watch and polysomnography, specifically acceleration and heart rate metrics, to develop a machine learning algorithm for distinguishing between sleep and wake states. This will be useful in determining exactly when a person has fallen asleep, leading to more accurate metrics for sleep time and quality, which will improve the overall health of the person.

The relevance of this research is highlighted by the growing body of literature that explores various methodologies for sleep analysis. One example is of such a study is done by Stuburić et al. (2020) who employed deep learning, utilising convolutional neural networks (CNN) and long-short term memory networks (LSTM) to detect sleep stages based on non-invasive signals, achieving an accuracy of 55%. This work was interesting, because of the fact that neural networks were used to analyse sleep. Unfortunately, the accuracy was not very high, hence this report will try different strategies. Another example, which is more similar to this report, was done by Smith et al. (2022). They extended the application of machine learning to sleep stage classification, achieving accuracies ranging from 74% to 96% using EEG and EMG data. Specifically, random forest and artificial neural network achieved notable accuracies of 96% and 93%, respectively. These values are quite promising, which shows that this technique has some merit. It would be interesting to analyse whether the accuracy for simply classifying between sleep and wake states would be even higher. Lastly, Sundararajan et al. (2021) addressed the need for cost-effective sleep measurement tools using random forest with wrist-worn accelerometers, achieving a 73.93% F1 score in classifying sleep states. This is quite high and seems like a working technique in classifying sleep stages.

Building upon this foundation, this report aims to analyse data from the Apple Watch. The dataset includes acceleration and heart rate metrics alongside gold-standard polysomnography sleep labels. The combination of the analysis of this dataset with

techniques similar to the ones in the existing literature will help the development of an accurate and practical machine learning algorithm for sleep-wake classification.

Problem Formulation

The research question that this report will answer is: Can machine learning algorithms effectively classify sleep and wake states using Apple Watch data, and how does the proposed approach compare to existing methodologies? The challenge is to develop a machine learning algorithm that can accurately classify sleep and wake states based on acceleration and heart rate data collected from the Apple Watch and polysomnography. The target audience for this report are healthcare professionals, researchers, and individuals interested in monitoring their sleep patterns. If successful, people could use this machine learning algorithm to assist them with their sleep monitoring, specifically by having an accurate measure of sleep time.

In this project feature values are heart rate and motion data from the Apple Watch, namely x-axis acceleration data. These were selected based on domain knowledge, because they are closely linked to falling asleep. For example, steps were dropped, because they are not as important in sleep, which would confuse the machine learning algorithms more than help with classification.

The label value is the labelled sleep, which originally was divided into five categories ranging from wake to REM sleep, but for this purpose it was divided into binary bins. 0 as wake state and values greater than 1 were simply classified as sleep state and were assigned the value of 1.

Dataset Description

The data is from the Apple Watch, photoplethysmography and polysomnography, collected by Olivia Walch (Walch, 2019) and it contains acceleration (in units of *g*), heart rate (bpm, measured from photoplethysmography), as well as labelled sleep scored from gold-standard polysomnography, as seen from Table 1. Data were collected at the University of Michigan from June 2017 to March 2019, and there are 31 subjects in total.

The data was cleaned by removing missing values and removing steps, since it was deemed unnecessary for sleep classification. Steps, y- and z-axis motion data were also removed, because people should not move except in the x-axis when they are falling asleep.

Table 1. Description of data

Variable	Type	Description
acc	Float	Motion (acceleration)
hrt	Integer	Heart rate (bpm)
stp	Integer	Steps (count)
stg	Integer	Labelled sleep (wake = 0, N1 = 1, N2 = 2, N3 = 3, REM = 5)

Observations

Group-level

In Figure 1, it can be seen that there is no clear and distinct division between states of being asleep and awake. This goes in line with the fact that people are very different. There are so many factors that influence sleep in such different ways, for example individual differences in sleep patterns, varying sleep onset times, and unique activity levels. This highlights the need for a modern approach in differentiating sleep and wake states, which is where machine learning algorithms, such as logistic regression and random forest, can help.

On the other hand, even without a clear division, we can see that there are clusters of sleep and wake data points within the dataset. One observation that can be made about these clusters is that most of the sleep data is with relatively low heart rate, when awake data is more spread out. The motion data does not seem to influence the wake or sleep states. These clusters indicate local patterns and trends that the machine learning algorithm could potentially find. While the individual variations make it more difficult to establish a universal threshold for sleep and wake classification, the algorithms could use these clusters to learn about sleep/wake states.

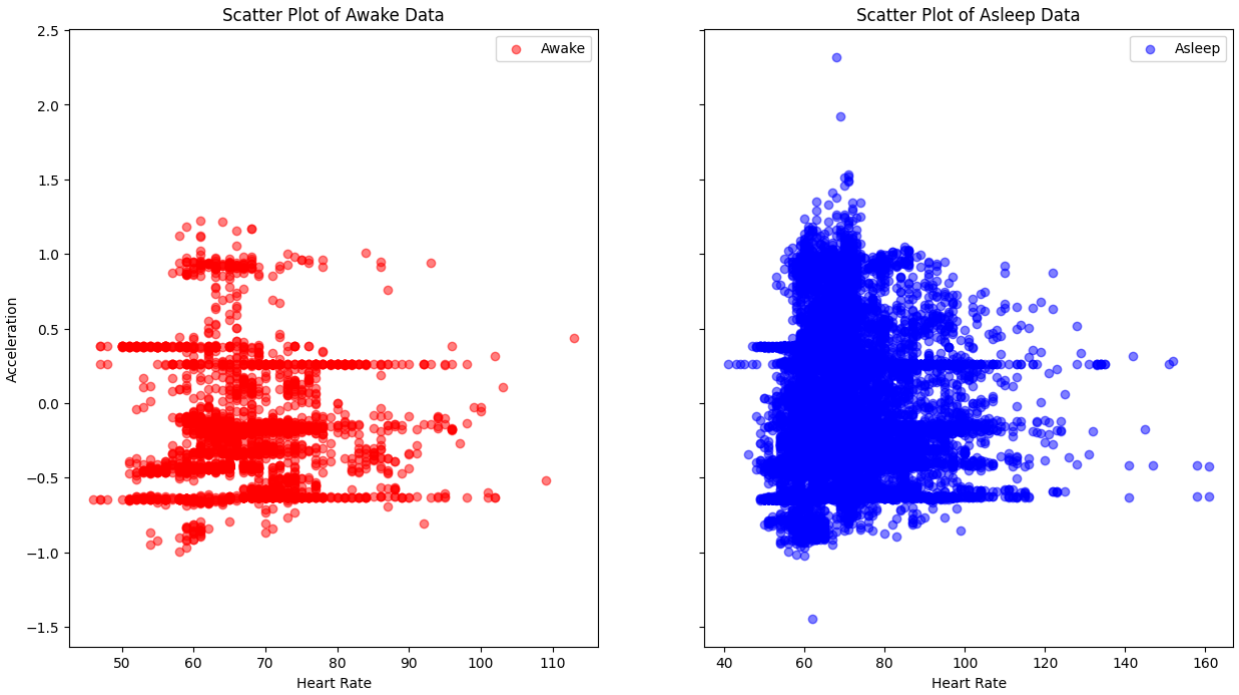

Figure 1. 3D Scatter Plot of data

Subject-level

When looking into the subject-level analysis, the focus shifts from the broader trends seen in group data to the unique patterns shown by an individual. A random subject has been chosen for a more in-depth examination of their sleep data. This could reveal some specific trends that may not be evident when examining averaged, general-level data.

The chosen subject's sleep data is quite similar to the general one, in the fact that there is no clear pattern. Notably, Figure 2. highlights a limited number of wake states.This could be attributed to the fact that the experiment initiated as the individuals were preparing to sleep. This would explain the abundance of sleep data with its different categories and a lack of wake data points. The distribution of sleep and wake states for this individual shows that on average the wake states occur when the acceleration data is between 0 and -0.5, while the sleep data occurs usually at zero. This observation aligns with the expectation that the subject moves more when they are awake than when they are sleeping. However, surprisingly, the heart rate data shows that even on the individual level, there can be huge differences in when the subject is asleep. They range from 50 to 130 bpm, which is surprising to say the least. Typically sleep would lower the heart rate and not vary this much. This could suggest a flaw in the data or data analysis.

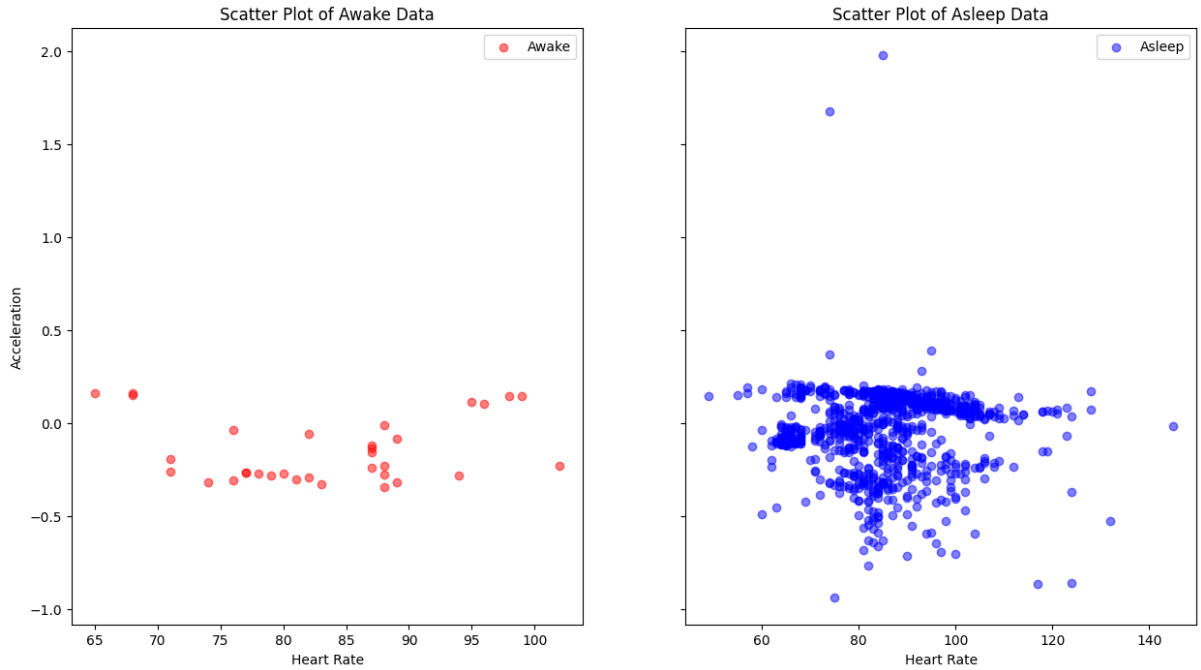

Figure 2. Individual sleep and wake, heart rate and acceleration graph

Conclusion for Observations

Based on the subject-level and group-level analysis, it is clear that there is no linear relationship within the data. The unique characteristics and variations shown by group-level and individual subjects highlight the fact that sleep patterns are very complicated. Despite this, the machine learning algorithms that are selected should be able to deal with non-linear data and classification, because they should be able to adapt to subtle patterns and variations.

Methods

Considering the complicated nature of the data as well as techniques from the existing literature like CNN, LSTM and random forest with their accuracy metrics it is clear that random forest was one of the most effective in classifying sleep stages. For that reason, random forest was the chosen method for this report. In addition to random forests, logistic regression was also used because it is well-suited to classify binary data. The metrics used to evaluate the model are accuracy, precision and the F1 score. In order to validate the results a training, testing split of 80-20 was employed.

Random Forest

Random forest is a machine learning algorithm that operates by constructing a number of decision trees during training and outputting the class that is the mode of the classes, for classification, or the mean prediction (regression) of the individual

trees. For this task, the type is classification, since we need to predict whether the subject is sleeping or awake. So the algorithm will take the mode of the classes.

The key principle behind a random forest is the aggregation of multiple decision trees to reduce overfitting and enhance predictive accuracy. Each tree in the forest is constructed using a random subset of the training data and a random subset of features at each split. This randomness makes the trees different, and the final prediction is determined by averaging, for regression, across all individual tree predictions.

Logistic Regression

Logistic Regression is a statistical method used for binary classification tasks in machine learning. It predicts the probability of an observation belonging to a particular class. It is particularly well-suited for problems where the dependent variable is binary, meaning it has only two possible outcomes, which is why it was chosen for this problem.

The fundamental idea behind logistic regression is to model the relationship between the independent variables and the probability of a particular outcome using the logistic function (sigmoid function). The logistic function transforms any input into a range between 0 and 1, making it suitable for representing probabilities. The logistic regression model calculates a weighted sum of the input features, and this sum is then passed through the logistic function to produce the predicted probability.

The logistic regression equation is given by:

$$P(Y = 1) = \frac{1}{1+e-(\beta_0+\beta_1 X_1+\beta_2 X_2+...+\beta_n X_n)} \qquad [1]$$

Results

After cleaning the data, both machine learning algorithms were implemented with some success, but it became clear that some improvements are needed. The logistic regression model showed a moderate level of success in predicting outcomes. While random forests model demonstrated a much higher accuracy. However, both models have specific areas that need improvement to improve their effectiveness.

Logistic Regression

The model showed an overall accuracy of 48%, which is not very good, but not terrible in predicting outcomes within the dataset. In the classification report,

precision for the wake state was found to be 12%, meaning that when the model predicted it, it was accurate only 12% of the time. On the other hand, precision for the sleep state stood at a more substantial 90%, showing a higher accuracy when predicting the sleep state. The recall for outcome wake and sleep states was 58% and 47%, respectively, meaning that the models' ability to get the actual instances of each outcome was limited. The F1-scores were 20% for wake and 61% for sleep, which suggests that the model is significantly better at identifying sleep states rather than wake states.

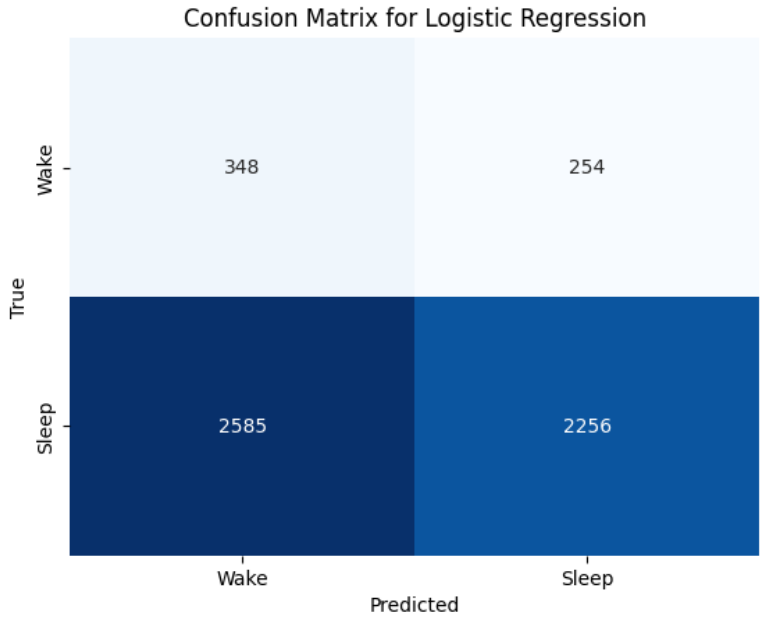

Figure 3. Confusion matrix for Logistic Regression

Analysing the confusion matrix in Figure 3. it supports the conclusion from the classification report. Namely, the model correctly identified 348 instances of true wake states but misclassified 254 instances as sleep states. Which is an approximate 50/50 split between correct and wrong guesses. Additionally, it accurately predicted 2585 instances of true sleep states but misclassified 2256 instances as wake states.

Random Forests

Random Forests model turned out to be a better model with an accuracy of 85%. Looking at the results more carefully, the precision for outcome wake state was 22%, which shows that when the model predicted wake state, it was correct 22% of the time. Conversely, precision for outcome sleep state exhibited a significantly higher 91%, meaning that there is a high accuracy when predicting sleep state. The recall for outcome wake state was 20% and for sleep state 92%, showing the model's skill in predicting the actual instances of outcome sleep state. The F1-scores were 21%

for wake state and 92% for sleep state, highlighting a robust performance in capturing both precision and recall for outcome sleep state.

Similarly to the logistic regression, when analysing the confusion matrix in Figure 4. it further supports the conclusions drawn from the classification report. The model correctly identified 100 instances of true wake states but misclassified 411 instances as sleep states. Additionally, it accurately predicted 4137 instances of true sleep states but misclassified 353 instances as wake states.

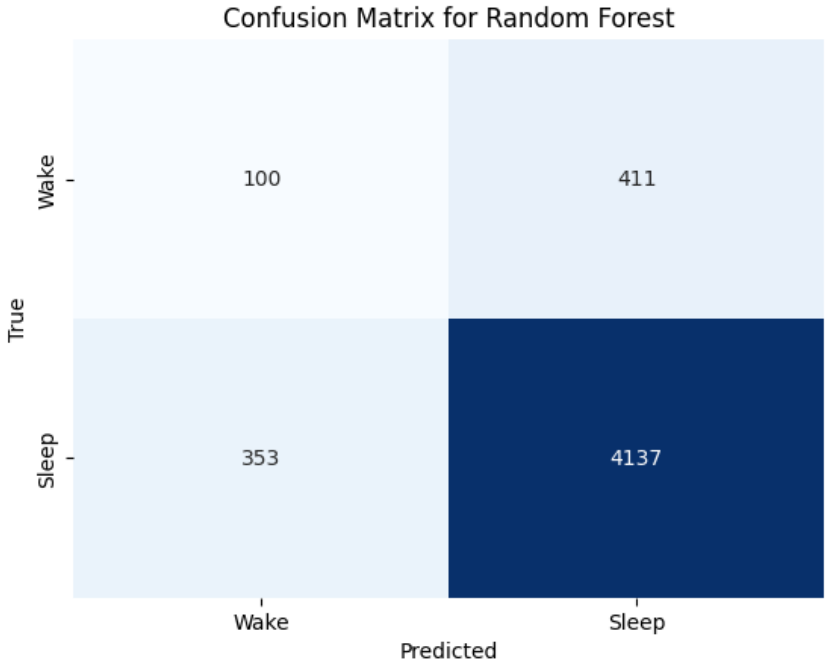

Figure 4. Confusion matrix for Random Forest

Discussion

The results obtained from logistic regression and random forests machine learning algorithms give some food for thought and provide an interesting lens through which we can analyse their performance and discuss potential improvement ideas. Even though generally, the results were not excellent, they still provide some interesting ideas for future work. Overall, the random forest classifier was the better model, which can be seen from Figure 5. It is better or equivalent in every metric, except recall of wake states. When it is better, it's almost twice as good as the logistic regression model. This supports the existing literature that pointed out that random forests was one of the best performing machine learning algorithms.

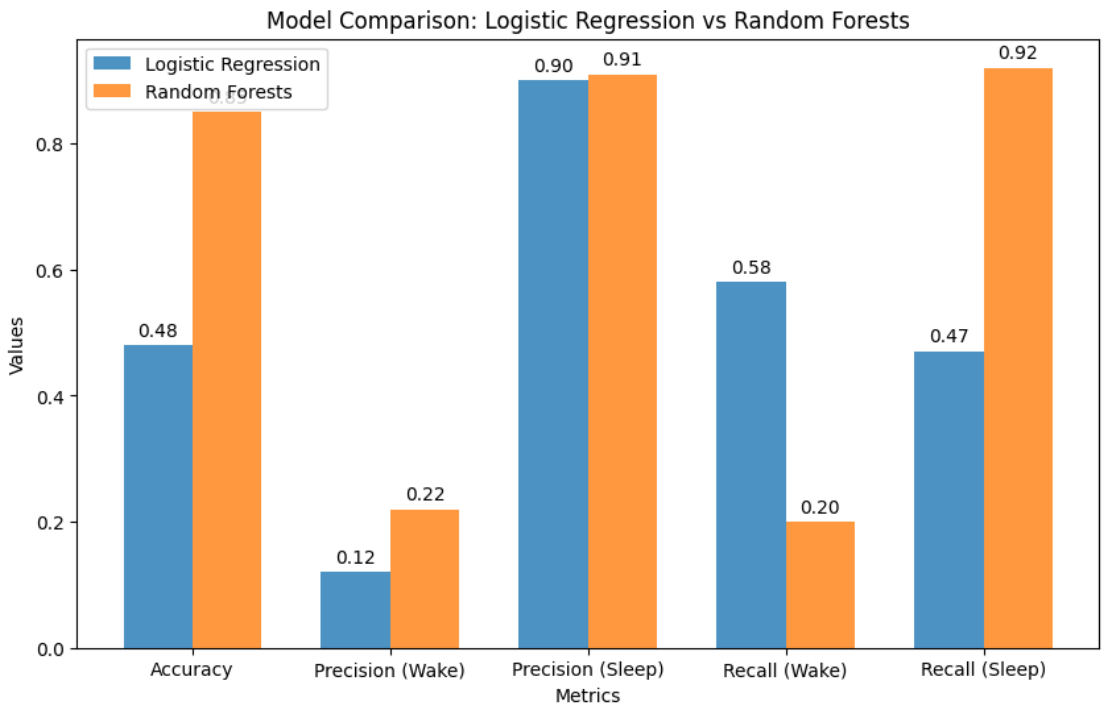

Figure 5. Model Comparison of Logistic Regression and Random Forests

There should also be some discussion about the data. Since the data was in separate folders, they had to be merged together and time matched before using machine learning models. Considering the seeming randomness of heart rate and motion data for each state, it could be because the data was not matched properly. This would have led to a completely random set of labels, accelerations and heart rates that did not match in real life and the models were trained on that. A possible improvement on this topic is about the data and making sure that the heart rate, motion and polysomnography results are matched in time.

Logistics Regression

The logistic regression model, characterised by its simplicity and interpretability, had a low measure of accuracy of 48%. The classification report made it clear that there were significant differences in precision, recall, and F1-scores between the two outcomes—wake and sleep states. One potential reason for this is the nature of logistic regression, which may not be able to capture the variety of subject-level sleep patterns effectively, even though it is good with binary classification. Because of this, the machine learning algorithm should not be disregarded completely, instead future research may consider using it for a single individual if they show low variation in data when falling asleep. A personalised logistic regression algorithm, trained specifically on an individual's data, could potentially enhance the model's predictive capabilities and be a useful asset in classifying between sleep states.

Additionally, the imbalanced distribution between sleep and wake states within the dataset might have contributed to the model's low performance. With fewer instances of wake states, the model may be biassed toward predicting sleep states, which leads to a higher accuracy in the more common outcome. Addressing these differences in the number of sleep and wake states within the dataset could also potentially improve the performance of logistic regression.

Random Forests

Unlike logistic regression, for the random forests model the classification report showed a significantly higher accuracy of 85%. In addition, it demonstrated better precision, recall, and F1-scores for both wake and sleep states. This was not a particularly surprising result, because the existing literature already proved that this machine learning approach is capable of displaying good results in predicting sleep/wake states. This could stem from the fact that random forests is able to capture the variation of patterns within the data, which leads to a better performance compared to logistic regression.

Examining the confusion matrix supported the idea that it could correctly predict sleep states. Random forests are better at handling imbalanced datasets, which could be one reason why it was better at handling this particular problem. However, there can also be some improvement, especially in accurately predicting instances of wake states. This is to be expected, because of the low number of wake states. One improvement suggestion is using a dataset with a balanced set of sleep and wake states by, for example, wearing the Apple Watch and PSG for eight hours during the day and night, so that there are approximately equal amounts of time of wake and sleep states.

Conclusion

To answer the question that was asked in the beginning: “Can machine learning algorithms effectively classify sleep and wake states using Apple Watch data, and how does the proposed approach compare to existing methodologies?” the question has to be divided into two parts. For the first part, this report applied two machine learning algorithms – logistic regression and random forests – which had limited success in classifying sleep and wake states. Logistic regression was shown to not be able to effectively classify sleep and wake states (accuracy 48%), while random forests seemingly could with an accuracy score 85%. The details from the classification reports and confusion matrices gave some insights into the strengths and weaknesses of each model. For the second part, the performances of the algorithms were weaker than in the existing literature. This could be because of the fact that the data was imbalanced and had complicated individual-level variance that was difficult to capture.

Logistic regression could be useful from an individual standpoint, as a personalised machine learning algorithm tailored to individual sleep patterns, but this has to be investigated in future research. On the other hand, the random forests model showed better performance, potentially, because it was able to effectively capture complicated patterns within the data. The model excelled, especially in predicting sleep states. Nevertheless, there remained areas for improvement for both models, for logistic regression the handling of imbalance datasets should be improved and for random forests, the improvement area lies in refining predictions related to wake states.

This report aids to the understanding of machine learning applications in sleep pattern prediction as well as highlights the need of tailored approaches and constant model refinement. Future work should explore logistic regression as a personalised machine learning strategy and random forests should be improved in wake state prediction.

References

Smith, A., Anand, H., Milosavljevic, S., Rentschler, K. M., Pocivavsek, A., & Valafar, H. (2022). Application of Machine Learning to Sleep Stage Classification. *2021 International Conference on Computational Science and Computational Intelligence (CSCI)*. https://doi.org/10.48550/arXiv.2111.03085

Stuburić, K., Gaiduk, M., & Seepold, R. (2020). A deep learning approach to detect sleep stages. *Procedia Computer Science*, *176*, 2764–2772. https://doi.org/10.1016/j.procs.2020.09.280

Sundararajan, K., Georgievska, S., Lindert, B. H., Gehrman, P. R., Ramautar, J., Mazzotti, D. R., Sabia, S., Weedon, M. N., van Someren, E. J., Ridder, L., Wang, J., & van Hees, V. T. (2021). Sleep classification from wrist-worn accelerometer data using random forests. *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-020-79217-x

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.

Walch, O., Huang, Y., Forger, D., & Goldstein, C. (2019). Sleep stage prediction with raw acceleration and photoplethysmography heart rate data derived from a consumer wearable device. *Sleep*, *42*(12). https://doi.org/10.1093/sleep/zsz180