Sleep & Awake Classification

Using Consumer Grade ENMO data

IFN646 Biomedical Data Science
Queensland University of Technology

Due date 10/11/20 11:59pm

Student 1: Nebojsa Ajdarevic – N10434348

Student 2: Karen Hau – N9445943

Student 3: Alex Conroy – N8324841

Introduction

Our body's demand for sleep has been somewhat of a puzzle for scientists for over millennia. However, there are some things we know for certain; sleep is of the utmost importance to a healthy and functioning life. The negative health impacts of a lack of sleep range from obesity, weight gain [1] and diabetes [2] to cardiovascular diseases [3] and various psychiatric problems [4]. As a result of this increased risk getting to understand sleep has been a target of research. Currently, the gold standard in sleep monitoring is the use of Polysomnography (PSG) where subjects would have to spend a night in a dedicated sleep laboratory while being hooked up to various machines. As expected, this is expensive to conduct and not at all pleasing for the subjects to experience. Currently, wearables are considered an alternative to this.

Clinical use of actigraphy to monitor the sleep-wake cycle has been plentiful, especially in the case of monitoring the normal sleep-wake cycles of participants. They have been proven time and time again to be useful and accurate tools for measuring the sleep-wake cycles of participants as compared to PSG [5], [6]. Clinical grade actigraphy watches like the Phillips Actiwatch uses an accelerometer to infer sleep-wake cycles using a classification algorithm. This is obviously cheaper than PSG and still useful. It does, however, have its drawbacks too, as it is still by no means cheap. Furthermore, there is a cheaper alternative, consumer-grade watches like the apple watch. The apple watch uses Euclidean Norm Minus One (ENMO) which also uses accelerometer data, this can be used to determine the sleep-wake cycle and has been done successfully in the past [7].

The widespread appeal of consumer-grade smartwatches has allowed them to become far more affordable than the clinical-grade watches like the Phillips Actiwatch. These watches are a potential goldmine of data as they have an enormous pool of users whose data could be accessed by researchers. Furthermore, these smartwatches provide many more data sources aside from the accelerometer data such as built-in magnetometer, gyroscope, and in some cases, heart rate sensors. Access to this data could provide benefits to a wide variety of research and comparing the accuracy of these watches to clinically proven ones is of the utmost importance for researchers.

A study has already been conducted wherein the validity of the apple watch ENMO data to predict sleep-wake cycles were tested against the Philips Actiwatch [7]. However, as this is a landmark study these results must be reproduced. However, the strategy opted with, in this report is to reproduce the Actiwatch activity counts data from the apple watch ENMO data. This permits the running of the activity counts classification algorithm on these converted values. Accurate classification using a consumer watch will open doors for cheaper research and big data sleep research in the future.

Based on previous research, it is hypothesised that consumer-grade smartwatches can be used for sleep-wake classification under normal sleeping conditions.

Methodology

The dataset was originally collected by Queensland University of Technology researchers from 14 participants, 9 males, 5 females, between April and May of 2018. Participants wore both clinical and commercial products for two consecutive nights in a home setting. The devices were both worn on their non-dominant hand. The initial data is missing one night of data due to user error, leaving 27 nights of data collected.

The dataset used by this report consists of processed data collected from an accelerometer in both devices. The clinical data consists of an activity count and classification provided by the equation shown previously. The consumer-grade data consists of a Euclidean Norm triaxial acceleration vector (ENMO), that was calculated using the below equation. These were both associated with a de-identified date-time stamp at 15-second epoch intervals.

$$ENMO(A) = \sqrt{(A^2{}_x + A^2{}_y + A^2{}_z)} - 1$$

This process is shown in figure 1, with changes occurring from the data originators methodology from Sleep Classification onwards.

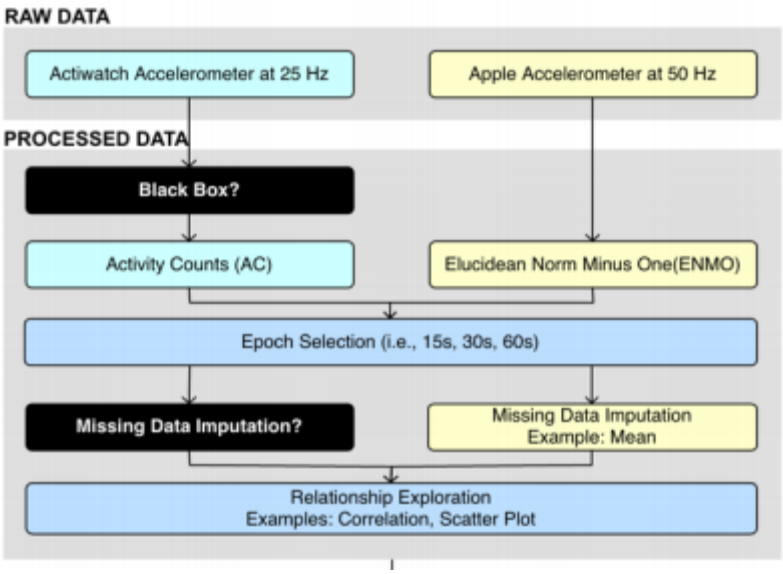

Figure 1 Original Data Processing Method

This process is shown in figure 1, with changes occurring from the data originators methodology from Sleep Classification onwards.

For this study, two approaches could be taken to produce a classification for comparison of the two types of hardware.

- Reproduce the activity count in such a way as to use the clinical-grade classification algorithm

- Produce a unique classification algorithm that can work directly with the consumer-grade data.

It was decided to use the same classification algorithm as the clinical-grade, with the hope that this would allow for a more accurate comparative analysis to be possible. This way any differences would be down to the nuances of the different grade data collection as opposed to issues with using two different classification algorithms.

The clinical grade algorithm computes the total activity counts using the weighted sum below, for a given epoch of e and activity count of a_e.

$$totalCounts(e) = 0.04\ \Sigma^{-5}{}_{i=-8} a_{e+i} + 0.2\ \Sigma^{-1}{}_{i=-4} a_{e+i} + 4a_e + 0.2\ \Sigma^{4}{}_{i=1} a_{e+i} + 0.04\ \Sigma^{8}{}_{i=5} a_{e+i}$$

This allows for the classification to be designated as awake or asleep given the values related to certain threshold values. With a sleep classification occurring if it is less than or equal to the given threshold, with a greater than occurrence resulting in an awake classification. The Initial thresholds used for this study were 20 and 40 as suggested by the Phillips Actiwatch guide.
After some initial data exploration, it appeared that linear regression was a suitable starting point for this dataset. Therefore, a linear regression analysis was conducted to alter the ENMO into a state that could be fed into the total counts' algorithm for classification, which was then compared to the original classification. It is noted that the linear regression was trained on the first 20 nights of sleep data.

After the linear regression was trained, we used the last 7 nights of data to test the model with the weighted sum Phillips algorithm as classification on this test dataset. This did the classification for us. The threshold of 40 performed better than 20 and was the one used here. Furthermore, as it is commonly practised when using actigraphy to look for the first 5 minutes of uninterrupted sleep, and set all the epochs before that to wake, and then the last 5 minutes of uninterrupted sleep, and also set all the epochs after that to wake.

The initial recall performance was initially an issue, it was surmised that there was a common error in the modelling performance. However, after the addition of the alteration to uninterrupted sleep classification mentioned in the paragraph above, there was a distinct uptick in the recall performance metric from ~0.25 to ~0.8. This is a significant improvement and closer to the 0.85 found in the original report. This now makes precision the worst-performing metric. Additional checks were performed to handle any outliers in the data, but these did not affect the overall classification accuracy.

Results

Results: Ordinary least squares

Model:	OLS	Adj. R-squared:	0.719
Dependent Variable:	Actiwatch activity counts	AIC:	281449.3880
Date:	2020-10-13 17:15	BIC:	281466.4570
No. Observations:	37590	Log-Likelihood:	-1.4072e+05
Df Model:	1	F-statistic:	9.624e+04
Df Residuals:	37588	Prob (F-statistic):	0.00
R-squared:	0.719	Scale:	104.52

	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
const	-2.3034	0.0554	-41.5749	0.0000	-2.4120	-2.1948
Apple Watch ENMO	1009.2220	3.2532	310.2273	0.0000	1002.8457	1015.5983

Omnibus:	64396.197	Durbin-Watson:	1.899
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2762056226.612
Skew:	-10.477	Prob(JB):	0.000
Kurtosis:	1330.797	Condition No.:	62

Figure 2 Linear Apple Watch Regression Model Output Summary

As per figure 2 above for every 1-point increase in the ENMO data the value for Actiwatch data increases by 1009.222 with a standard error of 3.25. However, as this model breaks the assumptions of linear regression, as can be seen in figure 3 below, it cannot be used to determine coefficients but rather, only for prediction. This linear regression has an adjusted R squared of 0.719 suggesting that the ENMO data accounts for roughly 72% of the variance observed within the Actiwatch activity counts data.

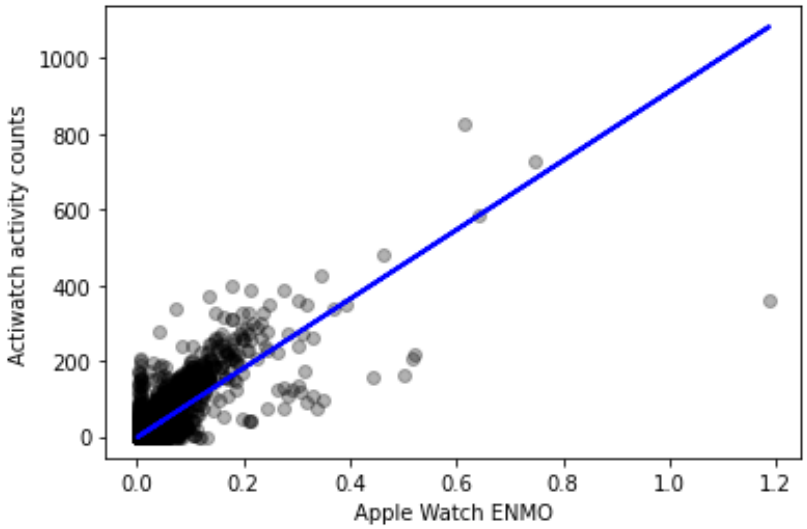

Figure 3 Apple ENMO vs Activity count for all candidates and nights

So, while it is clear from figure 4 below the assumptions for linear regression are broken. However, this model is used solely for prediction and not understanding the relationships between variables, therefore the model is still viable to be used in this way albeit with limitations.

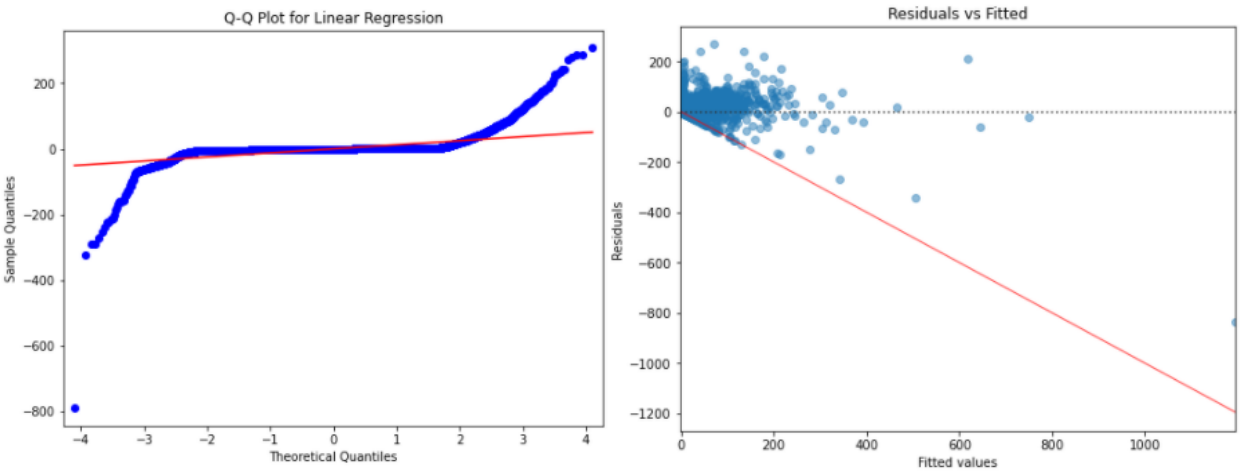

Figure 4 Linear Regression Assumptions

The performance metrics used to investigate the results of the model are defined in the below table. These are determined using the four outcomes of a classification problem: true positive (TP), true negative (TN), false positive (FP), and false-negative (FN).

Measure	Formula
Accuracy	(TP +TN)/(TP+TN+FN+FP)
Recall	TP/(TP+FN)
Specificity	TN/(TN+FP)
Precision	TP/(TP+FP)
F1 Score	2*Precision*Recall/ (Precision+Recall)

Table 1 Definition of Classification Performance Metrics

These metrics were then run in different test/train split conditions. Which investigated whether there was any significant improvements if the data was split using a traditional hold-out method versus removal of outliers, training on the first night of candidate data and testing on the second night of data, or randomisation.

	Accuracy	Precision	Recall	True Negative	F1 Score
1st 20 (train)	0.9640	0.8979	0.8350	0.9847	0.8653
No outliers	0.9569	0.8352	0.8586	0.9727	0.8467
First night (train)	0.9747	0.9338	0.7667	0.9948	0.8420
Random (20)	0.9653	0.9007	0.7948	0.9882	0.8444

Table 2 Comparison of Performance Metric Results for Different Training Approaches

These results shown in table 2 show that overall, there is no real improvement for any of the alternative methods compared to the hold-out. Therefore, it was determined to use the standard hold-out method with an approximate 75/25 train/test split.

After using simple linear regression to convert ENMO data into Actiwear data it was evident that the data needed to be modelled better despite the high accuracy of the conversion. The accuracy of the final model was 96% recall of 83% precision 90%, and F1 87%. This is evident in table 3 below.

Measure	Performance
Accuracy	96.40%
Recall	83.50%
Specificity	98.47%
Precision	89.79%
F1 Score	86.53%

Table 3 Overall Classification Performance for Activewear Data

The confusion matrix in figure 5 below shows a succinct summary of the results reported in table 3 above, where 0 represents asleep and 1 represents awake.

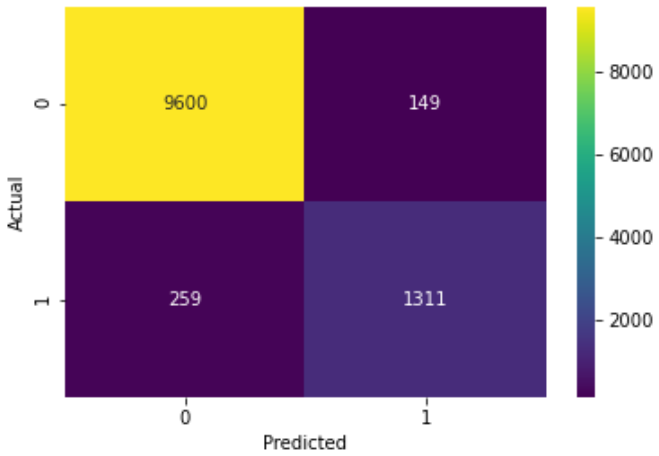

Figure 5 Regression Model Confusion Matrix

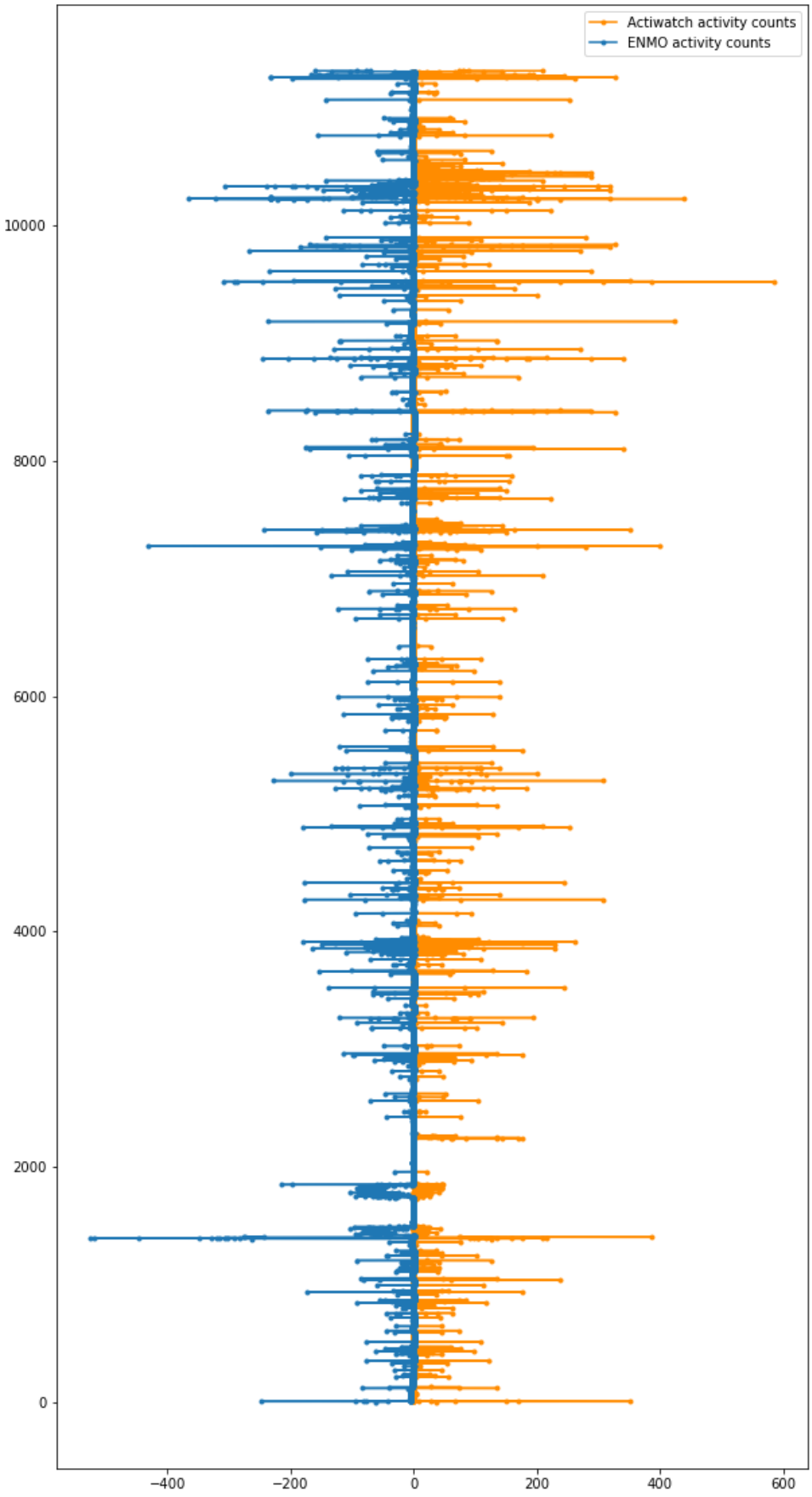

Figure 6 Every night's measurement at 15 sec epochs ENMO in blue (left) and Activity Counts in orange (right)

Figure 6 above shows that there are some errors within the conversion, but mostly the valleys and hills match each other. To improve the model, the data were modified by dealing with the outliers. As a result, the linear algorithm performed better, however, this did not have any substantial impact on the results as such we used the original unaltered linear regression as above in figure 6.

Discussion

The results suggest that linear regression is sufficient at converting Apple watch series 1 ENMO data into Phillips Actiwatch activity count data for a range of normal sleepers. These results are as expected given this dataset and the exploratory data analysis. This supports the hypothesis that consumer-grade smartwatches can be used for sleep-wake classification under normal sleeping conditions.

This model, as with all others, comes with its own set of limitations. As this data was collected mainly during the times closest to sleep this can provide many potential generalisation limitations to the model we produced. For instance, this model would fail at accurately predicting the sleep-wake cycle during the daytime as it may consider someone sitting down or not moving much as asleep. This is due to the inherent imbalance within the dataset. Furthermore, while the validity of the Activewear watch has been tested it is not perfect. Our model assumes that it is the ground truth, as a result, this conversion has two degrees of separation from reality and is compounding on the errors underlying the Actiwatch. Tuning to predict the classification of the Actiwatch classifier may be inherently steering away from reality.

Furthermore, with the limitations in the size and depth of the dataset, it is unlikely that the results produced from the ENMO data will be applicable in a generalized setting, i.e. for health-related sleep studies. This is due largely to the model's inability to accurately capture inter-night waking episodes. Due to the relatively few instances given in the dataset for the model to learn from. This is one of our recommendations for future works, as the focus of this report was on the reproducibility of commercial-grade hardware in a clinical setting. Rather than the production of a fully capable model to be utilised should this hold. As this is not something that is tested within the dataset and will require exploration before the ENMO data classification can be considered for deployment for use in any real-world settings.

Conclusion

The report shows that the Gen 1 Apple watch ENMO data is reasonably capable of predicting sleep awake classification when compared to the Philips Actiwatch. It shows that there is a linear relationship between Apple watch ENMO data and Philips Actiwatch activity count data. Additionally, after conversion, the Philips Actiwatch algorithm is a reasonable approach for classifying this data again as it receives high accuracy specificity, precision, and recall. This gives favourable support to the hypothesis that consumer-grade smartwatches can be used for sleep-wake classification under normal sleeping conditions.

While the prediction accuracy indicated was quite favourable, several improvements can be recommended as topics for future analysis. Including analysis on abnormal sleeping data, tests between Apple ENMO and higher-order clinical equipment.

Data availability

The original QUT Research Data can be found at https://doi.org/10.25912/ 5cc28f62e81ad16. As stated by the authors, the data can be accessed as a zip file containing 27 CSV files, one for each night recorded. The data contains four columns: a timestamp, Actiwatch activity count, Actiwear classification, and Apple watch ENMO data. The date found in timestamp has been modified to maintain participant privacy, with the times themselves preserved to maintain the 15-sec epoch. The data has been made available under the Creative Commons Attribution 4.0 International license (CC-BY 4.0) terms.

The modifications made to this data for this report can be found using the code in the appendix.

References

[1] M. Watanabe, H. Kikuchi, K. Tanaka, and M. Takahashi, 'Association of short sleep duration with weight gain and obesity at 1-year follow-up: a large-scale prospective study', *Sleep*, vol. 33, no. 2, pp. 161–167, Feb. 2010, doi: 10.1093/sleep/33.2.161.

[2] K. Spiegel, K. Knutson, R. Leproult, E. Tasali, and E. Van Cauter, 'Sleep loss: a novel risk factor for insulin resistance and Type 2 diabetes', *J Appl Physiol (1985)*, vol. 99, no. 5, pp. 2008–2019, Nov. 2005, doi: 10.1152/japplphysiol.00660.2005

[3] E. Kasasbeh, D. S. Chi, and G. Krishnaswamy, 'Inflammatory aspects of sleep apnea and their cardiovascular consequences', *South Med J*, vol. 99, no. 1, pp. 58–67; quiz 68–69, 81, Jan. 2006, doi: 10.1097/01.smj.0000197705.99639.50.

[4] E. J. Paavonen *et al.*, 'Sleep problems of school-aged children: a complementary view', *Acta Paediatr*, vol. 89, no. 2, pp. 223–228, Feb. 2000, doi: 10.1080/080352500750028870.

[5] M. Shin, P. Swan, and C. M. Chow, 'The validity of Actiwatch2 and SenseWear armband compared against polysomnography at different ambient temperature conditions', *Sleep Sci*, vol. 8, no. 1, pp. 9–15, Mar. 2015, doi: 10.1016/j.slsci.2015.02.003.

[6] M. de Zambotti, F. C. Baker, and I. M. Colrain, 'Validation of Sleep-Tracking Technology Compared with Polysomnography in Adolescents', *Sleep*, vol. 38, no. 9, pp. 1461–1468, Sep. 2015, doi: 10.5665/sleep.4990.

[7] S. Roomkham, M. Hittle, J. Cheung, D. Lovell, E. Mignot, and D. Perrin, 'Sleep monitoring with the Apple Watch: comparison to a clinically validated actigraph', *F1000Res*, vol. 8, p. 754, May 2019, doi: 10.12688/f1000research.19020.1.