Discrimination of sleep and wake periods from a hip-worn raw acceleration sensor using recurrent neural networks

Marc Weitz[1,*], Shaheen Syed[1], Laila A. Hopstock[2], Bente Morseth[3], Dilip K. Prasad[1], and Alexander Horsch[1]

[1]Department of Computer Science, Faculty of Science, UiT The Arctic University of Norway, Tromsø, Norway
[2]Department of Community Medicine, Faculty of Health Sciences, UiT The Arctic University of Norway, Tromsø, Norway
[3]School of Sport Sciences, Faculty of Health Sciences, UiT The Arctic University of Norway, Tromsø, Norway
[*]marc.weitz@uit.no

ABSTRACT

The use of accelerometers has become an established method in population research. Accelerometers are small body-worn sensors that can monitor movement patterns and cycles over several days under free-living conditions. A key requirement for any accelerometer-based analysis is the reliable discrimination of sleep and wake episodes. However, many studies using hip-worn sensors either instruct the participants to remove the sensor over night or use algorithms developed for wrist placement which are known to perform poorly on hip data. Here we present a new algorithm to differentiate sleep from awake time in raw hip acceleration data using a machine learning approach. Validated on sleep estimates from a subset of the Tromsø Study, the proposed algorithm outperformed the standard algorithms by discriminating sleep and wake correctly in 93.81% (95% CI: 0.72) of the time (F1 score = 0.93, 95% CI: 0.01). The value of our algorithm lies in the improved recognition of sleep and movement in large studies where manual scoring is unfeasible due to the amount of data. Even though the algorithm was developed to learn characteristic hip movements, it can easily be adapted for other sensor placements or study populations.

Introduction

The use of accelerometers has become a popular assessment method to monitor human physical activity. Accelerometers are small body-worn sensors that can monitor subjects for a week or more without intervention. Due to its non-invasive and unobtrusive application, accelerometers are used in many population and clinical studies. The monitoring of sleep and wake periods by accelerometers to estimate sleep-related parameters is frequently called actigraphy[1]. Actigraphy is increasingly used as it has the advantage of providing objective estimates of sleep habits under free-living conditions[2,3]. Moreover, it is considerably cheaper than the gold-standard method polysomnography, which usually requires the participant to stay the night in a sleep laboratory. This has made actigraphy also a feasible objective assessment method in many free-living cohort studies.

Since the start in the early nineties, several algorithms have been proposed to effectively recognize sleep patterns from acceleration[4–8]. Acceleration, in this context, is most often expressed in counts per minute (cpm), a threshold-based procedure that measures the number of times the raw acceleration signal exceeds a certain threshold in the given period (epoch). This data aggregation strategy has been developed to overcome the data storage problems of that time. However, device specific count estimations make it hard to compare results from different studies and devices. Even though analyzing count data is still common, newer methods are most often developed for raw acceleration data. Raw acceleration data, as defined by multiples of the earth's gravitation g ($1\,\mathrm{g} = 9.81\,\mathrm{m\,s^{-2}}$) is known to give the researcher more freedom in data analysis and to increase the consistency between studies. It also provides much richer data and increases transparency of the analysis.

Even though raw acceleration data is increasingly used[9], only a few algorithms have been published that utilize the high resolution of the raw signal to detect sleep[6–8]. Previously published algorithms either detect sleep with[6] or without[7] a sleep diary from wrist-worn accelerometers by using wrist-acceleration specific features. Even though it has been argued that these methods could be applied to acceleration data collected from other wear locations, there is currently no scientific literature that supports their interpretation[10].

To the best of our knowledge, no algorithm has yet been presented for hip-worn accelerometers. Hip-worn acceleration is mostly used in physical activity research as it has been argued that hip movement provides a better estimation of daily energy expenditure[11]. The removal of sleep episodes is important in this research for closer examination of the subject's physical activity. Some studies asked the participants to remove the accelerometer over night[12,13], likely to avoid confusion of sleep and

sedentary behavior. For that reason, a method that simultaneously provides reliable energy expenditure and sleep estimates without requiring the subject to wear multiple sensors is highly desired.

One reason that no such method exists today might be that algorithms for hip-worn sensors are lacking. To overcome this problem, we used different machine learning models to learn the underlying characteristic movement patterns. Machine learning, and neural networks in particular, are known to be able to model complex relationships without the need of manual feature engineering. Previous studies that have used machine learning in actigraphy have either focused on count data[14] or incorporated wrist-specific features[8]. Due to the richer signal, raw acceleration data or commonly used frequency features are even better candidates for machine learning based pattern recognition than count data.

The goal of this study was to develop an algorithm to efficiently and reliably differentiate sleep from awake time in raw hip acceleration data. The general idea behind our algorithm is to learn the characteristic movements of sleep using neural networks. To come up with the most reliable algorithm, we compared different machine learning models and preprocessing strategies. The most promising machine learning candidate is the Long Short-Term Memory network (LSTM) due to its ability to capture temporal dependencies by recurrent connections within the network[15]. To test different degrees of general and temporal complexity, we additionally selected the following types of machine learning models: generalized linear model (GLM), standard multi-layer artificial neural network (MLP) and multi-layer Elman recurrent neural network (RNN). As previous algorithms published on raw acceleration data raise serious theoretical concerns in the application on hip-worn data as mentioned earlier, baseline comparisons are conducted against the well established Cole-Kripke[4] and Sadeh[5] algorithms. These two algorithms are also the standard algorithms of ActiGraph's own data analysis software ActiLife. Besides raw data, we tested the machine learning models also on the original ActiGraph counts and manually engineered frequency features.

Methods

Study population

The data of 444 participants of the Tromsø Study were included in this study. The Tromsø Study is a population-based cohort study in the municipality of Tromsø in Norway that includes seven data collection waves conducted between 1974 and 2016[16,17]. The dataset described in the following was acquired in the seventh wave of the Tromsø Study (Tromsø7) from 2015-16. Tromsø7 was approved by the Regional Committee for Medical Research Ethics (REC North ref. 2014/940) and the Norwegian Data Protection Authority, and all participants gave written informed consent. The usage of data in this study has been approved by the Data Publication Committee of the Tromsø Study. Furthermore, all methods were carried out in accordance with relevant guidelines and regulations (i.e. Declaration of Helsinki).

Among the 444 participants that were included in this study, 252 were female and 192 male, with a median age of 63.5 ($IQR = 13.0$). The participants had a mean height of 169 cm (SD = 8), a mean weight of 77 kg (SD = 14) and a mean body mass index of 27 $\mathrm{kg\,m^{-2}}$ (SD = 4).

Instrumentation

We used data from two sensors that the participants wore simultaneously, the Actiwave Cardio (CamNtech Ltd, Cambridge, UK) and the ActiGraph wGT3X-BT (ActiGraph, Pensacola, FL). The Actiwave Cardio is a miniature single-lead ECG, acceleration and inclination recorder with a dynamic range of 8 g (1 g = 9.81 $\mathrm{m\,s^{-2}}$) that was worn on the participant's chest for approximately 24 hours. Acceleration was sampled along three axes (vertical, mediolateral and anteroposterior) at a sampling rate of 32 Hz. The ActiGraph only recorded triaxial acceleration, but with a higher sampling frequency of 100 Hz and a longer recording time of seven consecutive days on the right hip. From the Actiwave Cardio's raw triaxial acceleration data and the subjects' inclination, ground truth labels were derived based on the manufacturer's manual by M.W. and S.S. independently. Both annotators were instructed to identify the time the subject went to bed (characterized by a drastic change in the activity pattern from activity to barely no activity and a position change from an active to a lying position) and the time the subject got out of bed again (characterized by a longer preceding episode of low activity followed by a higher amount of activity and a change of position from a lying to an active position). To ensure data quality, only subjects with an intersection over the union (IoU) score of over .9 between the two annotations were included in this study. For all subjects meeting this requirement, the intersection of the two annotators' rating was labelled as sleep and aligned with the corresponding data of the ActiGraph sensor.

Experimental Design

To develop our algorithm, we compared several machine learning models (GLM, MLP, RNN and LSTM) and inputs (counts per minute, raw acceleration and frequency features) and compared them against baseline algorithms (Cole-Kripke and Sadeh). For the machine learning models with hyperparameters (MLP, RNN and LSTM), hyperparameter searches were conducted in order to find the best performing models.

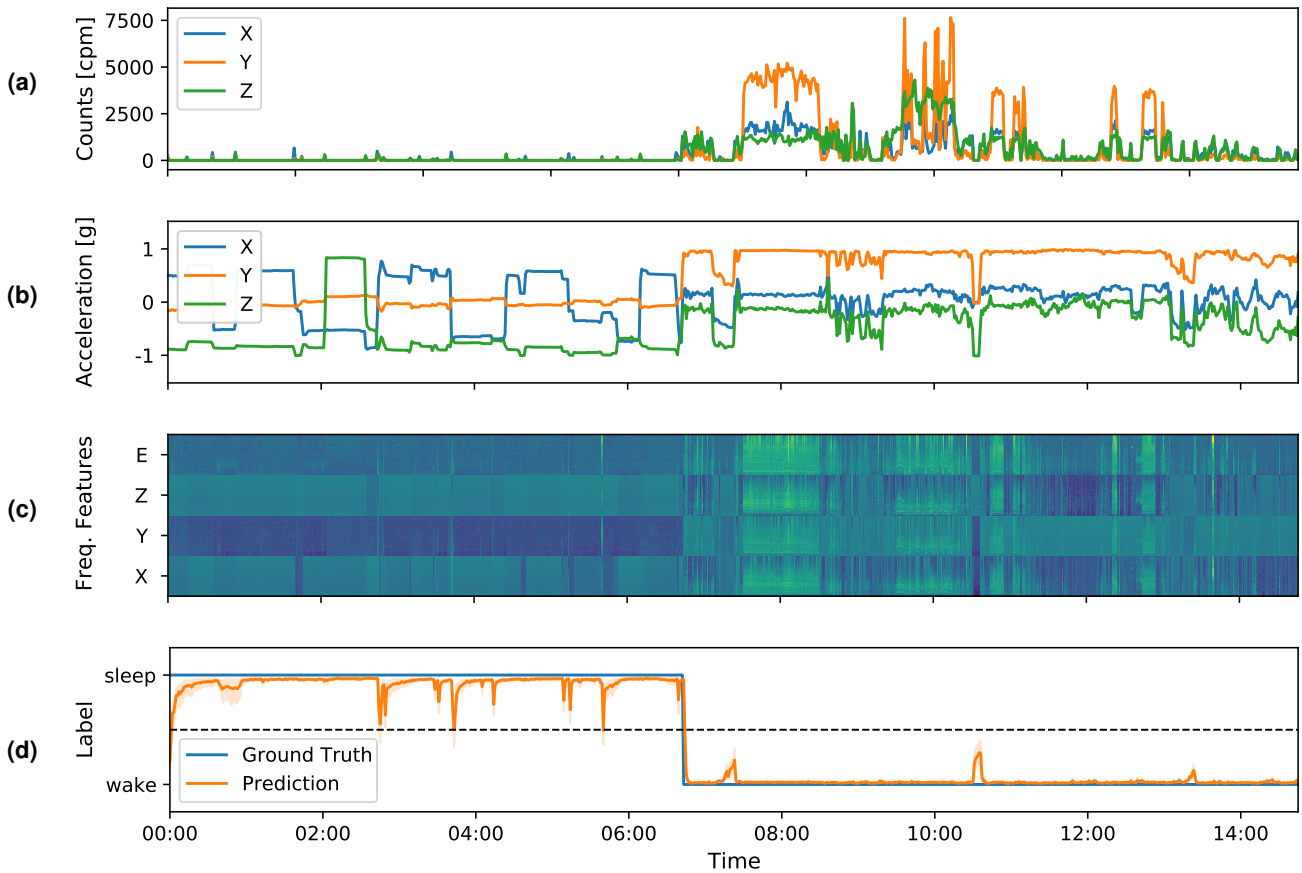

Figure 1. Visualization of input types, labels and predictions for one participant. (a) Counts per minute for all three axes (b) Raw triaxial acceleration. (c) Calculated frequency features for the triaxial raw data and their Euclidean Norm minus one. (d) Ground truth labels and the final algorithm's predictions with 95% CI based on 10-fold cross-validation.

Input and feature engineering

Counts per minute

Count data refers to an aggregated type of acceleration data. The general idea behind counts is the number of times the acceleration signal exceeds a certain threshold during a specific epoch length. Counts are sensitive to the manufacturer and the exact algorithms are unknown. A recently published study reported to have reverse engineered the ActiGraph algorithm by a combination of various filters[18]. We used the original ActiGraph counts obtained by the ActiLife software and temporally aligned them with the other data. In this study an epoch length of 1-minute was used. A visualization of the y-axis count data is shown in Figure 1a.

Raw acceleration

By raw acceleration, we refer to the original triaxial signal measured by the accelerometer in gravitational units g ($1\,\mathrm{g} = 9.81\,\mathrm{m\,s^{-2}}$). The data we used were originally sampled at 100 Hz, but we aggregated them to 1-minute time resolution by averaging each minute's sample values (see Figure 1b). This was done because recurrent neural networks are only capable of modelling time dependencies in terms of data points. Using 100 Hz data would have yielded an effectively learned lag under 1 s.

Frequency features

Besides count and raw data, we evaluated the models also on manually engineered features. Mel filter banks are common features in many speech recognition applications[19]. The idea behind these features is to decompose the signal over a certain episode into different frequency bands. As in speech processing considerably higher sampling rates are common, we applied the procedure on non-overlapping one minute windows with 6000 data points each. Additionally to each of the three axes, we

calculated frequency features for the Euclidean Norm Minus One (ENMO):

$$ENMO_t = \sqrt{x_t^2 + y_t^2 + z_t^2} - 1 \tag{1}$$

The ENMO is a commonly used measure to represent the acceleration independently from the direction as the euclidean norm represents the lengths of the triaxial acceleration vector. One is subtracted to account for the general gravitation on earth.

For each of the non-overlapping windows, a Hamming window was applied and the power spectrum was calculated for the weighted window w by

$$P(w_i) = \frac{|FFT(w_i)|^2}{N} \tag{2}$$

with FFT being the Fast Fourier Transform and N the number of points to use. The window function is usually applied to account for overlap, but we kept it to stick with general methodology. In the final step, triangular filters were applied on the power spectrum to obtain frequency bands (see Figure 1c for a visualization of the filter bank energies). We used 512 points for the Fast Fourier Transform and 40 triangular filters per channel.

Models and algorithms

Traditional approaches

To estimate baseline performance, we applied the Sadeh algorithm[5] and the Cole-Kripke algorithm[4]. Both algorithms are also implemented in ActiGraph's own data analysis software ActiLife and provide a comparison to the current standard procedure. They both use the y-axis over a certain window, only (11-minute and 7-minute window, respectively). We used the original published parameters and did not further optimize them. To provide comparability with the other methods, we evaluated the algorithms on the same ten splits as the other models.

Machine learning models

We optimized and compared the performance of four common machine learning models: GLM, MLP, RNN and LSTM. As the most common baseline model, we implemented a GLM to discriminate between sleep and wake for each time step. Additionally, we trained several MLPs with varying depth (1, 2, 4 layers) and hidden layer size (1, 2, 4, 8, 16, 32, 64 neurons) on the same task. Compared to the GLM, MLPs are able to model considerably more complex relationships, but they do not take into account temporal information. Each layer of the MLPs consisted of a fully connected linear layer succeeded by a dropout and a Rectified Linear Unit (ReLU) layer, except for the last layer which consisted only of a fully connected linear layer with sigmoid activation. To additionally allow for temporal dependencies, we trained Elman RNNs and LSTMs. We trained the RNNs and LSTMs with varying depth (1, 2, 4 layers) and hidden state size (1 ,2 ,4, 8, 16, 32, 64 dimensions). Dropout with $p = .5$ was applied after all recurrent layers except for the final layer[20]. From all trained models, we chose the hyperparameters that led to the highest F1 score. In case of a tie, we chose the model with less parameters.

Training procedure

The data was normalized with the mean and standard deviation of the entire data set. We applied 10-fold cross-validation. Within each fold, we split the data into a train and validation set on a 80:20 ratio. Each fold was trained until the loss of the validation set did not further decrease. We used stochastic gradient descent and the binary cross entropy loss function to train the parameters of all machine learning models. A batch size of 8 and a learning rate of .7 based on prior experiments was applied. Models and code were developed with Pytorch[21] and scikit-learn[22] in Python 3[23]. All code for the experiments can be found in the supplementary material.

The algorithms were evaluated in terms of accuracy, F1 score, precision, and recall with specific focus on the F1 score as the harmonic mean of precision and recall. The Cole-Kripke and the Sadeh algorithms were only evaluated on count data, whereas the four machine learning models GLM, MLP, RNN and LSTM were evaluated on count, raw and frequency data. A hyperparameter search on depth (number of layers) and hidden state size (for RNN and LSTM) and hidden layer size (for MLP) was performed to determine the best-scoring machine learning models. The GLM had no hyperparameters to optimize.

Results

LSTMs achieved the highest F1 scores and accuracies for all input type conditions (see Table 1). The highest F1 score and accuracy were achieved when using the manually engineered frequency features, achieving a 94% accuracy of correctly labelling one minute of data at a F1 score of 0.93. The traditional algorithms, Sadeh and Cole-Kripke, were among the worst-performing methods, both on accuracy and F1 score, at a high recall, though. Among all data points in our data set that were associated with sleep, the Cole-Kripke identified 98% and the Sadeh algorithm 96% correctly as sleep. However, from

Table 1. Results for the best methods per input type. Shown are average metrics over 10-fold cross-validation with the 95% confidence interval given in brackets. Cole-Kripke and Sadeh are the standard methods used by the device manufacturer. For count, raw and frequency data, a hyperparameter search was done for each model and the best performing model with respect to the F1 score was selected. If two models tied, the model with less parameters was chosen.

Input	Method	Accuracy	F1 Score	Precision	Recall
Counts per minute	Cole-Kripke	.69 (±.00)	.74 (±.01)	.60 (±.01)	.98 (±.00)
Counts per minute	Sadeh	.77 (±.00)	.79 (±.01)	.68 (±.01)	.96 (±.00)
Counts per minute	GLM	.73 (±.01)	.76 (±.02)	.64 (±.03)	.95 (±.00)
Counts per minute	MLP	.74 (±.01)	.77 (±.02)	.66 (±.03)	.94 (±.00)
Counts per minute	RNN	.77 (±.02)	.79 (±.02)	.69 (±.03)	.94 (±.01)
Counts per minute	LSTM	.91 (±.01)	.90 (±.02)	.89 (±.03)	.92 (±.01)
Raw acceleration	GLM	.89 (±.01)	.88 (±.02)	.87 (±.02)	.92 (±.01)
Raw acceleration	MLP	.89 (±.01)	.89 (±.02)	.87 (±.02)	.93 (±.01)
Raw acceleration	RNN	.91 (±.01)	.91 (±.01)	.89 (±.02)	.95 (±.01)
Raw acceleration	LSTM	.92 (±.01)	.92 (±.01)	.91 (±.01)	.94 (±.01)
Frequency features	GLM	.87 (±.01)	.85 (±.02)	.86 (±.02)	.86 (±.02)
Frequency features	MLP	.89 (±.01)	.88 (±.02)	.86 (±.02)	.92 (±.01)
Frequency features	RNN	.92 (±.01)	.91 (±.01)	.90 (±.01)	.93 (±.01)
Frequency features	LSTM	.94 (±.01)	.93 (±.01)	.93 (±.01)	.94 (±.01)

all predicted sleep minutes, only 60% and 67%, respectively, were labelled correctly. Recall was generally high amongst all models and algorithms. Precision, on the other hand, varied between the different inputs and was worst for count data. Details of the hyperparameter search for the MLP, RNN and LSTM models and the selected hyperparameter configurations can be found in the supplementary material Table S1.

Besides accuracy and F1 score, we also evaluated the best performing model's ability to estimate total sleep time (TST), defined as the sum of all minutes labeled as sleep over the day. This model, a one layer LSTM with a hidden state of 4 trained on the frequency features, underestimated the TST on average by 10.8 minutes. We used a modified Bland-Altman plot[24] to investigate the agreement of the model's prediction with the ground truth. Instead of the average of the two measures, we plotted the difference against the ground truth TST. As shown in Figure 2a, no systematic prediction error can be observed. In general, the algorithm seems to resemble the distribution of TST in the data set well (Figure 2b). However, a few outliers seem to exist for which the model either estimated all or no time as sleep (upper- and lowermost points in the left, and left- and rightmost bars in the right plot).

Discussion

The aim of this study was to fill a gap in the processing of raw acceleration data. While raw acceleration data processing is increasingly used in many research fields, sleep algorithms have been dominantly proposed for wrist-worn sensors[6,7,10]. One reason for this is that it has been argued that wrist-worn sensors are better in measuring sleep patterns[1]. However, in particular in population research, hip-worn sensors are popular due to their advantage of more accurately estimating daily energy expenditure[11]. Manually annotating the data is cumbersome and not feasible for large cohorts or sample sizes, especially if recorded over multiple days. With several non-wear detection[25,26] and activity intensity prediction[27–31] algorithms in place, a reliable sleep detection algorithm was highly desired to effectively process raw acceleration data collected from hip-worn sensors.

Here, we presented a new algorithm to discriminate sleep and wake periods from raw data acquired by hip-worn accelerometers. The backbone of our algorithm is a LSTM network which learned the characteristic movement patterns of sleep and wake phases to accurately classify them. In particular, the richer raw signal leads to a better performance. This is in accordance with previous studies highlighting the benefits of raw acceleration data compared with count data[9,32]. In addition, we showed that the use of frequency features, which is common in speech recognition, can further enhance performance. The method can be used to automatically annotate raw hip acceleration of variable length and is independent of the accelerometer brand. When evaluating it on estimating the total sleep time, the algorithm underestimated the true value by only a few minutes.

Previous studies already mentioned performance issues of traditional algorithms to detect wake episodes from hip-worn sensors[33,34]. This observation was confirmed in our experiments as indicated by the combination of poor accuracy and high recall for the Cole-Kripke and Sadeh algorithm (Table 1). Both methods were originally developed for wrist-worn sensors and

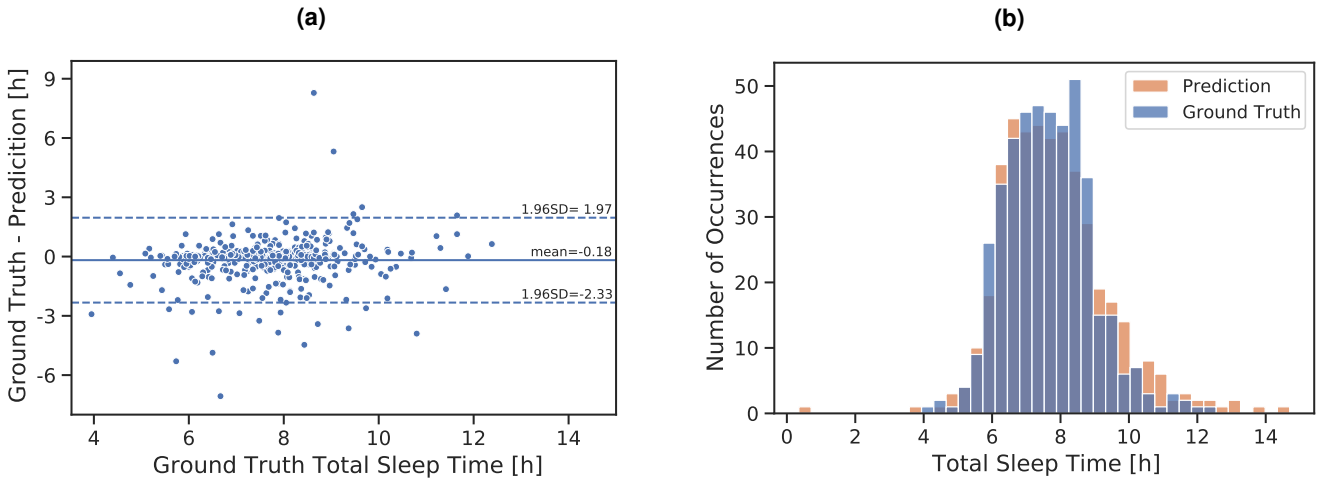

Figure 2. Comparison of the final algorithm's predictions against ground truth. (a) Modified Bland-Altman plot with 95% limits of agreement for the total sleep time estimates relative to the ground truth. (b) Frequency distributions of ground truth and prediction of total sleep time. Except of a few outliers, the proposed method shows high agreement with the ground truth data independent of the true total sleep time.

only later, due to a lack of methodology, increasingly used for hip data. All methods were good in identifying sleep episodes, but the machine learning models were better in detecting wake episodes, resulting in a higher precision, accuracy and F1 score. A high recall might be a minor problem in sleep research settings when the data mainly consists of sleep episodes or when only a suggestion for expert assessment is desired. However, under free-living long-term observation, where awake episodes are usually accounting for two-thirds of the data and no additional processing is favored, accurate estimates are of high importance.

The observed low precision for the Cole-Kripke and Sadeh algorithm might result from using counts as input data. As shown in Table 1, precision was generally low among all models and methods that used count data, suggesting that it is more difficult to differentiate between sleep and sedentary behaviour using counts, only. In particular, the Sadeh and the Cole-Kripke algorithm were developed for analysing only the y-axis data[4,5] while the other models were trained on the count data of all three axes. Using triaxial data as input seems to be particularly advantageous when using raw and frequency data. One explanation for this is that the direction of the acceleration vector can then be utilized. This might be particularly beneficial during episodes of low activity, when the acceleration vector approximately resembles the direction towards the earth's core. For hip-worn sensors, which are fixed with respect to the body's torso, this can ease posture estimation and most likely facilitates the discrimination between sleep (usually a laying position) and sedentary behaviour.

The use of frequency features seem to further enhance classification. The frequency features were derived by decomposing and filtering the signal to a low dimensional feature vector describing each minute. The resulting feature vectors keep a significantly higher amount of information than the averaging procedure applied to the raw data. Data aggregation in general was necessary as LSTMs are only capable of learning temporal dependencies of a certain amount of data points. Too high sampling rates can, for that reason, lead to poorer performance. The use of frequency features such as Mel filter banks or Mel Frequency Cepstral Coefficients (MFCCs) has become a popular technique to tackle the same problem in speech recognition where sampling rates are usually even higher. The use of frequency decomposition only received marginal attention in acceleration processing[35]. However, our study clearly highlights its benefits. Future research should consider and further explore the use of this technique.

A possible limitation of our study might be the use of an expert criterion derived from the Actiwave Cardio sensor for training and validation. Even though expert criteria have been used previously in sleep algorithm development[36], the established gold standard for measuring sleep is polysomnography. The used labels were derived based on the manufacturer's manual and, thus, reflect common manual accelerometer-based sleep scoring, but are potentially biased by the annotators. In any case, the proposed algorithm contributes to overcome the cumbersome and often unfeasible work of manually scoring subjects in large-scale free-living multi-day recording settings. Nevertheless, further validation against polysomnography is highly desirable.

The proposed method might also show some limitations when applied to new data. Even though all models were evaluated ten times on new previously unseen data, all data was sampled from the same population, namely mostly white adults living in the municipality of Tromsø in Northern Norway. Demographic biases are a well-known problem in both machine learning and

population research. A potential bias when applied on other study populations can, thus, not be precluded. As the method aims to model the temporal pattern between the sensor's measurement and the subject's sleep status, the generalizability to other data depends on the variation of sleep movements across study populations. This might, for instance, affect the scoring of subjects suffering various sleep diseases for which sleep assessment with previous methods is already known to be more difficult[37]. Also other factors such as age and sex might influence sleep movements and for that reason affect scoring performance. More research across different populations and more diverse data sets will be required to further investigate potential biases.

Conclusion

Here we presented the first algorithm to discriminate sleep and wake episodes from raw hip-worn accelerometers. The main application domain of the proposed method is in cohort studies in which sensor placement on the hip is often preferred over the wrist due to the better estimation of daily energy expenditure. The lack of such a method has previously led either to protocols instructing the participants to remove the sensor over night, or to the application of methods that were originally developed for the wrist and are known for performing poorly on hip acceleration. By training a neural network to learn the underlying movement patterns of sleep and wake time, we were able to classify sleep and wake episodes at an accuracy of 94% compared with an independently assessed expert criterion. While trained on hip acceleration, the proposed method can easily be adapted to other sensor placements or study populations by retraining the network. The final algorithm can automatically annotate raw acceleration data of variable length and is for this reason independent of the accelerometer brand. This will facilitate population research and other fields in which accelerometers are used, in particular when manual annotation becomes unfeasible due to the amount of data. To support the application and validation of the algorithm, a Python implementation is provided to enable researchers to obtain an annotated acceleration time series from an ActiGraph gt3x file with just a few lines of code.

References

1. Martin, J. L. & Hakim, A. D. Wrist Actigraphy. *Chest* **139**, 1514–1527, DOI: 10.1378/chest.10-1872 (2011).
2. Sadeh, A. The role and validity of actigraphy in sleep medicine: An update. *Sleep Medicine Rev.* **15**, 259–267, DOI: 10.1016/j.smrv.2010.10.001 (2011).
3. Smith, M. T. *et al.* Use of actigraphy for the evaluation of sleep disorders and circadian rhythm sleep-wake disorders: An american academy of sleep medicine systematic review, meta-analysis, and grade assessment. *J. Clin. Sleep Medicine* **14**, 1209–1230, DOI: 10.5664/jcsm.7228 (2018).
4. Cole, R. J., Kripke, D. F., Gruen, W., Mullaney, D. J. & Gillin, J. C. Automatic sleep/wake identification from wrist activity. *Sleep* **15**, 461–469, DOI: 10.1093/sleep/15.5.461 (1992).
5. Sadeh, A., Sharkey, M. & Carskadon, M. A. Activity-based sleep-wake identification: An empirical test of methodological issues. *Sleep* **17**, 201–207, DOI: 10.1093/sleep/17.3.201 (1994).
6. van Hees, V. T. *et al.* A Novel, Open Access Method to Assess Sleep Duration Using a Wrist-Worn Accelerometer. *PLOS ONE* **10**, e0142533, DOI: 10.1371/journal.pone.0142533 (2015).
7. van Hees, V. T. *et al.* Estimating sleep parameters using an accelerometer without sleep diary. *Sci. Reports* **8**, 12975, DOI: 10.1038/s41598-018-31266-z (2018).
8. Sundararajan, K. *et al.* Sleep classification from wrist-worn accelerometer data using random forests. *Sci. Reports* **11**, 24, DOI: 10.1038/s41598-020-79217-x (2021).
9. van Hees, V. T. *et al.* Challenges and Opportunities for Harmonizing Research Methodology: Raw Accelerometry. *Methods Inf. Medicine* **55**, 525–532, DOI: 10.3414/ME15-05-0013 (2016). Publisher: Schattauer GmbH.
10. Migueles, J. H., Rowlands, A. V., Huber, F., Sabia, S. & van Hees, V. T. GGIR: A Research Community–Driven Open Source R Package for Generating Physical Activity and Sleep Outcomes From Multi-Day Raw Accelerometer Data. *J. for Meas. Phys. Behav.* **2**, 188–196, DOI: 10.1123/jmpb.2018-0063 (2019).
11. Troiano, R. P. *et al.* Physical Activity in the United States Measured by Accelerometer. *Medicine & Sci. Sports & Exerc.* **40**, 181–188, DOI: 10.1249/mss.0b013e31815a51b3 (2008).
12. Mitchell, J. A. *et al.* A Prospective Study of Sedentary Behavior in a Large Cohort of Youth. *Medicine & Sci. Sports & Exerc.* **44**, 1081–1087, DOI: 10.1249/MSS.0b013e3182446c65 (2012).
13. Lee, I.-M. *et al.* Accelerometer-Measured Physical Activity and Sedentary Behavior in Relation to All-Cause Mortality. *Circulation* **137**, 203–205, DOI: 10.1161/CIRCULATIONAHA.117.031300 (2018).

14. Palotti, J. *et al.* Benchmark on a large cohort for sleep-wake classification with machine learning techniques. *npj Digit. Medicine* **2**, DOI: 10.1038/s41746-019-0126-9 (2019).

15. Hochreiter, S. & Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **9**, 1735–1780, DOI: 10.1162/neco.1997.9.8.1735 (1997).

16. Jacobsen, B. K., Eggen, A. E., Mathiesen, E. B., Wilsgaard, T. & Njølstad, I. Cohort profile: The tromso study. *Int. J. Epidemiol.* **41**, 961–967, DOI: 10.1093/ije/dyr049 (2011).

17. Sagelv, E. H. *et al.* Physical activity levels in adults and elderly from triaxial and uniaxial accelerometry. the tromsø study. *PLOS ONE* **14**, e0225670, DOI: 10.1371/journal.pone.0225670 (2019).

18. Brønd, J. C., Andersen, L. B. & Arvidsson, D. Generating ActiGraph counts from raw acceleration recorded by an alternative monitor. *Medicine & Sci. Sports & Exerc.* **49**, 2351–2360, DOI: 10.1249/mss.0000000000001344 (2017).

19. Purwins, H. *et al.* Deep Learning for Audio Signal Processing. *IEEE J. Sel. Top. Signal Process.* **13**, 206–219, DOI: 10.1109/JSTSP.2019.2908700 (2019).

20. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).

21. Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. *Adv. neural information processing systems* **32**, 8026–8037 (2019).

22. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

23. Van Rossum, G. & Drake, F. L. *Python 3 Reference Manual* (CreateSpace, Scotts Valley, CA, 2009).

24. Martin Bland, J. & Altman, D. G. Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet* **327**, 307–310, DOI: 10.1016/S0140-6736(86)90837-8 (1986).

25. Syed, S., Morseth, B., Hopstock, L. A. & Horsch, A. Evaluating the performance of raw and epoch non-wear algorithms using multiple accelerometers and electrocardiogram recordings. *Sci. Reports* **10**, 5866, DOI: 10.1038/s41598-020-62821-2 (2020).

26. Syed, S., Morseth, B., Hopstock, L. A. & Horsch, A. A novel algorithm to detect non-wear time from raw accelerometer data using deep convolutional neural networks. *Sci. Reports* **11**, 8832, DOI: 10.1038/s41598-021-87757-z (2021).

27. Hildebrand, M., van Hees, V. T., Hansen, B. H. & Ekelund, U. Age group comparability of raw accelerometer output from wrist- and hip-worn monitors. *Medicine science sports exercise* **46**, 1816–1824, DOI: 10.1249/mss.0000000000000289 (2014).

28. Hildebrand, M., Hansen, B. H., van Hees, V. T. & Ekelund, U. Evaluation of raw acceleration sedentary thresholds in children and adults. *Scand. J. Medicine & Sci. Sports* **27**, 1814–1823, DOI: 10.1111/sms.12795 (2017).

29. Hurter, L. *et al.* Establishing Raw Acceleration Thresholds to Classify Sedentary and Stationary Behaviour in Children. *Children* **5**, 172, DOI: 10.3390/children5120172 (2018).

30. Sanders, G. J. *et al.* Evaluation of wrist and hip sedentary behaviour and moderate-to-vigorous physical activity raw acceleration cutpoints in older adults. *J. Sports Sci.* **37**, 1270–1279, DOI: 10.1080/02640414.2018.1555904 (2019).

31. Crotti, M. *et al.* Development of raw acceleration cut-points for wrist and hip accelerometers to assess sedentary behaviour and physical activity in 5–7-year-old children. *J. Sports Sci.* **38**, 1036–1045, DOI: 10.1080/02640414.2020.1740469 (2020).

32. van Hees, V. T., Pias, M., Taherian, S., Ekelund, U. & Brage, S. A method to compare new and traditional accelerometry data in physical activity monitoring. In *2010 IEEE International Symposium on "A World of Wireless, Mobile and Multimedia Networks" (WoWMoM)*, 1–6, DOI: 10.1109/WOWMOM.2010.5534986 (2010).

33. Slater, J. A. *et al.* Assessing sleep using hip and wrist actigraphy: Hip and wrist actigraphy. *Sleep Biol. Rhythm.* **13**, 172–180, DOI: 10.1111/sbr.12103 (2015).

34. Zinkhan, M. *et al.* Agreement of different methods for assessing sleep characteristics: a comparison of two actigraphs, wrist and hip placement, and self-report with polysomnography. *Sleep Medicine* **15**, 1107–1114, DOI: 10.1016/j.sleep.2014.04.015 (2014).

35. Wohlfahrt, P. *et al.* Transitions in effective scaling behavior of accelerometric time series across sleep and wake. *EPL (Europhysics Lett.* **103**, 68002, DOI: 10.1209/0295-5075/103/68002 (2013).

36. Tudor-Locke, C., Barreira, T. V., Schuna, J. M., Mire, E. F. & Katzmarzyk, P. T. Fully automated waist-worn accelerometer algorithm for detecting children's sleep-period time separate from 24-h physical activity or sedentary behaviors. *Appl. Physiol. Nutr. Metab.* **39**, 53–57, DOI: 10.1139/apnm-2013-0173 (2014).

37. Conley, S. *et al.* Agreement between actigraphic and polysomnographic measures of sleep in adults with and without chronic conditions: A systematic review and meta-analysis. *Sleep Medicine Rev.* **46**, 151–160, DOI: 10.1016/j.smrv.2019.05.001 (2019).

Additional Information

Acknowledgement

This work was supported by the High North Population Studies at UiT The Arctic University of Norway.

Author contributions

M.W. conceived the idea. B.M., L.H. and A.H. collected the data. S.S. and M.W. annotated the data. M.W. designed and performed the experiments, derived the models and analyzed the data. M.W. wrote the manuscript. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Data availability

The data used in this study is available on request to the Tromsø Study Data and Publication Committee. Manual annotations can be made available upon request to the authors. All code used in this study is publicly available in the supporting repository at `https://github.com/trybnetic/tu7-acceleration-sleep-wake-classification`. The final algorithm is implemented in our Physical Activity Analysis Toolbox (PAAT) which is publicly available at `https://github.com/Trybnetic/paat`.

Supplementary Material

Table S1. F1 scores for all combinations of the performed hyperparameter search over the number of layers and hidden sizes/neurons of the hidden layer(s). Best scoring combinations that were further evaluated and reported in Table 1 are highlighted in bold.

Counts per minute

Layers							
				MLP			
1	.54 (±.38)	.53 (±.37)	.76 (±.06)	.76 (±.06)	.76 (±.07)	.76 (±.06)	.76 (±.07)
2	.00 (±.00)	.24 (±.39)	.71 (±.25)	**.77 (±.06)**	.77 (±.06)	.76 (±.07)	.76 (±.07)
4	.00 (±.00)	.00 (±.00)	.31 (±.40)	.60 (±.32)	.77 (±.06)	.77 (±.06)	.77 (±.06)
				RNN			
1	.77 (±.07)	.77 (±.07)	.69 (±.25)	**.79 (±.07)**	.62 (±.33)	.78 (±.09)	.72 (±.16)
2	.38 (±.41)	.61 (±.32)	.78 (±.08)	.64 (±.28)	.70 (±.25)	.65 (±.28)	.66 (±.25)
4	.00 (±.00)	.33 (±.43)	.63 (±.33)	.57 (±.34)	.53 (±.31)	.66 (±.27)	.72 (±.26)
				LSTM			
1	.78 (±.28)	**.90 (±.06)**	.84 (±.12)	.87 (±.06)	.85 (±.05)	.81 (±.08)	.84 (±.05)
2	.51 (±.44)	.52 (±.45)	.76 (±.27)	.77 (±.28)	.81 (±.09)	.85 (±.08)	.85 (±.05)
4	.09 (±.28)	.00 (±.00)	.05 (±.16)	.07 (±.23)	.09 (±.27)	.58 (±.40)	.77 (±.27)
Hidden Size	1	2	4	8	16	32	64

Raw acceleration

Layers							
				MLP			
1	.73 (±.29)	.79 (±.28)	**.89 (±.05)**	.89 (±.05)	.89 (±.05)	.89 (±.05)	.89 (±.05)
2	.16 (±.34)	.85 (±.14)	.89 (±.05)	.89 (±.05)	.89 (±.05)	.90 (±.05)	.89 (±.05)
4	.00 (±.00)	.00 (±.00)	.22 (±.30)	.89 (±.05)	.89 (±.05)	.89 (±.06)	.89 (±.05)
				RNN			
1	.88 (±.05)	.90 (±.05)	.89 (±.05)	**.91 (±.05)**	.91 (±.05)	.91 (±.05)	.89 (±.10)
2	.51 (±.45)	.87 (±.11)	.90 (±.05)	.90 (±.05)	.90 (±.05)	.90 (±.06)	.89 (±.07)
4	.09 (±.28)	.81 (±.29)	.89 (±.05)	.90 (±.05)	.81 (±.29)	.90 (±.04)	.74 (±.30)
				LSTM			
1	.91 (±.05)	**.92 (±.04)**	.92 (±.04)	.92 (±.04)	.91 (±.05)	.92 (±.05)	.91 (±.05)
2	.81 (±.29)	.90 (±.04)	.91 (±.04)	.90 (±.04)	.91 (±.05)	.91 (±.04)	.91 (±.05)
4	.00 (±.00)	.08 (±.25)	.17 (±.35)	.74 (±.39)	.81 (±.29)	.90 (±.04)	.90 (±.05)
Hidden Size	1	2	4	8	16	32	64

Frequency features

Layers							
				MLP			
1	.85 (±.05)	.87 (±.05)	**.88 (±.05)**	.87 (±.05)	.87 (±.05)	.87 (±.05)	.87 (±.05)
2	.33 (±.43)	.86 (±.06)	.87 (±.05)	.87 (±.05)	.87 (±.05)	.87 (±.05)	.87 (±.05)
4	.00 (±.00)	.00 (±.00)	.82 (±.10)	.87 (±.05)	.87 (±.05)	.88 (±.05)	.87 (±.05)
				RNN			
1	.88 (±.05)	.88 (±.05)	.89 (±.05)	.89 (±.04)	.89 (±.06)	**.91 (±.03)**	.90 (±.04)
2	.80 (±.28)	.90 (±.04)	.91 (±.03)	.90 (±.03)	.91 (±.04)	.90 (±.06)	.90 (±.05)
4	.35 (±.45)	.80 (±.28)	.89 (±.05)	.90 (±.05)	.90 (±.05)	.91 (±.04)	.90 (±.05)
				LSTM			
1	.90 (±.05)	.92 (±.03)	**.93 (±.04)**	.92 (±.03)	.93 (±.02)	.92 (±.04)	.93 (±.04)
2	.61 (±.43)	.82 (±.26)	.92 (±.04)	.91 (±.05)	.92 (±.05)	.93 (±.02)	.92 (±.02)
4	.00 (±.00)	.00 (±.00)	.53 (±.46)	.93 (±.04)	.92 (±.03)	.91 (±.04)	.91 (±.04)
Hidden Size	1	2	4	8	16	32	64

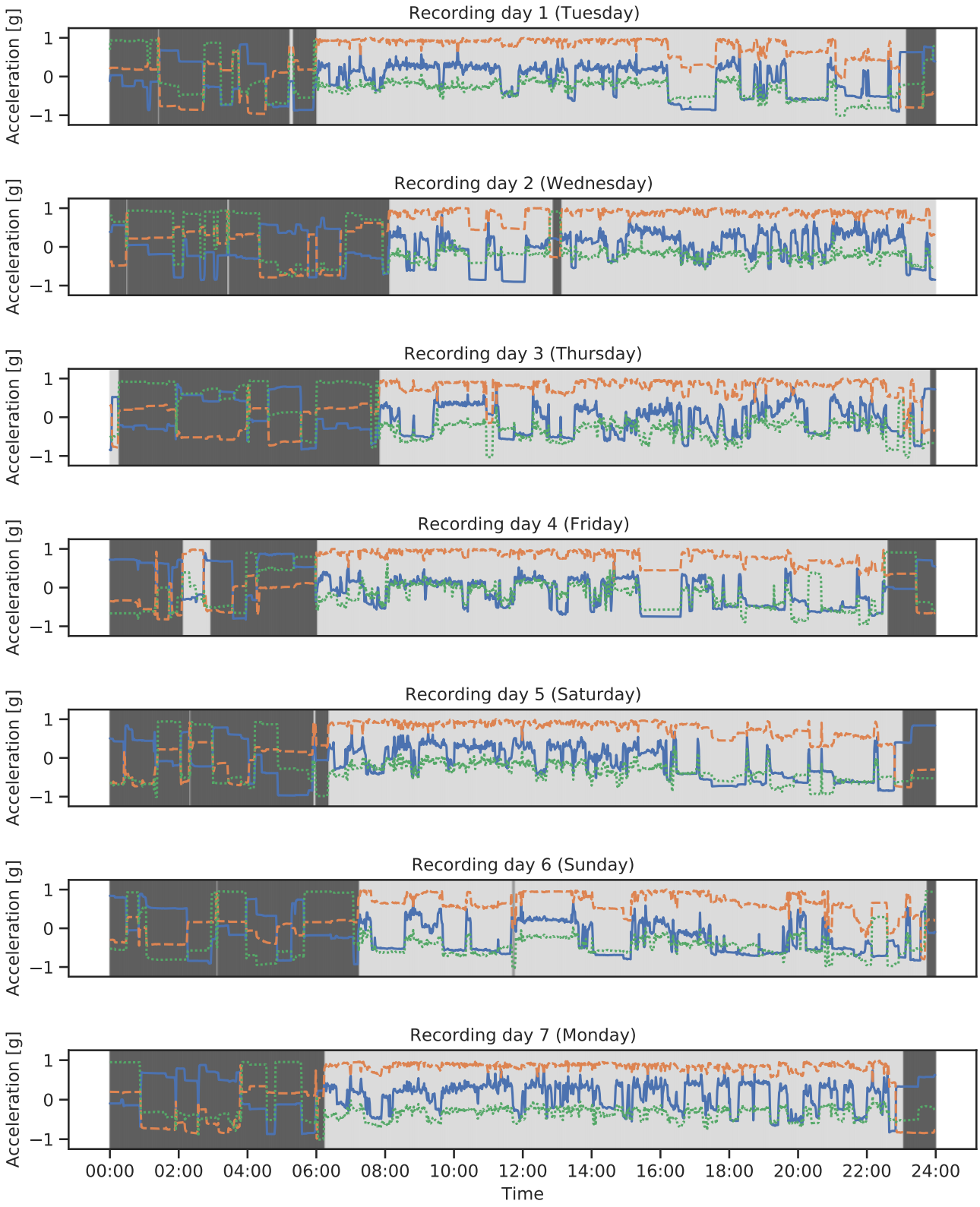

Figure S1. The proposed method can be used to automatically annotate acceleration time series of independent length. Here, one example application is shown for annotating the seven day recording of one participant of the Tromsø7 Study by sleep (dark grey), awake (light grey) and non-wear time (grey). The annotated data can then be used for further analysis.

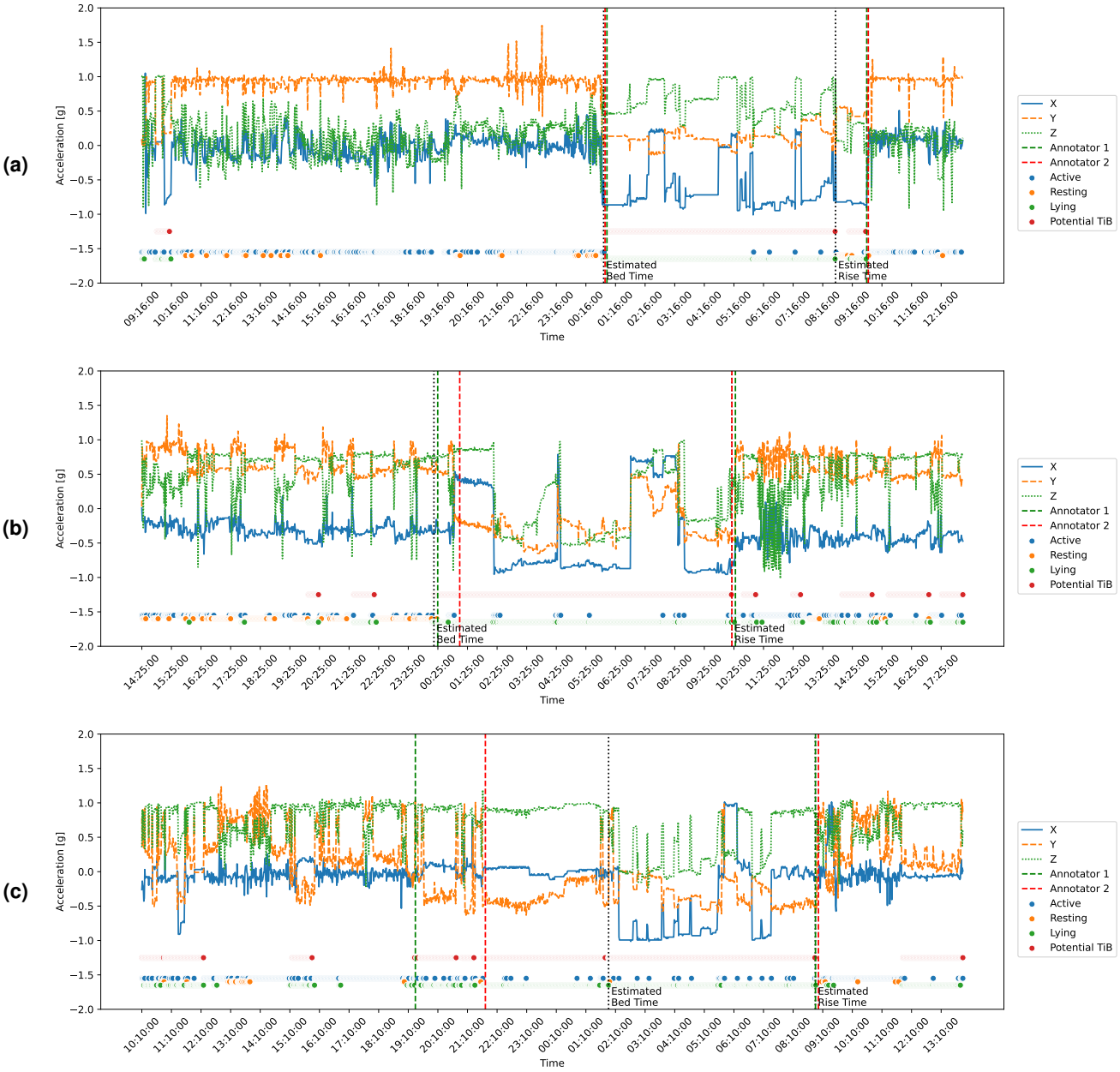

Figure S2. Examples from the annotation tool that was used to obtain the sleep labels from the chest-placed ActiWave sensor. (a) Both annotators corrected the estimated rise time with a very similar estimate. (b) Both annotators corrected the estimated bed time, but estimates vary slightly. (c) Both annotators corrected the estimated bed time, but estimates vary significantly. The bed time (characterized by a drastic change in the activity pattern from activity to barely no activity and a position change from an active to a lying position) and the rise time (characterized by a longer preceding episode of low activity followed by a higher amount of activity and a change of position from a lying to an active position) were determined independently by two annotators on one minute resolution. Estimates were proposed based on a lying position over a longer period, but checked or corrected by the annotators.

Table S2. Estimates for total time awake, non wear time (based on the Syed algorithm[26]) and total sleep time derived from the annotated data in Figure S1 per day and on average in minutes to give an idea of potential use.

Day of recording	0	1	2	3	4	5	6	Mean	SD
Total time awake	1035	940	976	1043	1005	986	1009	999.14	35.47
Total non wear time	0	0	0	0	0	5	0	0.71	1.89
Total sleep time	405	500	464	397	435	449	431	440.14	35.21