2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Influence of Study Composition on the Efficacy of Sleep Detection Using Actigraphy

Kevin Chao*, Bryan Fry, Kuldeep Singh Rajput, and Nandakumar Selvaraj

***Abstract*— Wearable actigraphy sensors have been useful tools for unobtrusive monitoring of sleep. The influence of the composition and characteristics of study groups such as normal sleep versus sleep disorders affecting the efficacy of sleep assessment using actigraphy has not been fully examined. In this study, we present multi-variate sleep models using actigraphy features obtained from wrist-worn sensors and evaluate the efficacy of sleep detection compared to the overnight polysomnography from two unique datasets: overnight actigraphy recordings in a control population of young healthy individuals (n=31) and 24-hour actigraphy recordings in a more heterogeneous population (n=27) comprised of normal and abnormal sleepers. We evaluate the performance of actigraphy derived logistic regression (LR) and random forest (RF) sleep models for both intra-dataset and inter-dataset training and cross-validation. Both the LR and RF sleep models for the healthy sleep dataset show an area under the receiver operating characteristic (AUROC) of 0.85±0.02 in the control sleep dataset among 50 random splits of training and testing evaluations. We find the AUROC performance from the heterogeneous sleep dataset involving sleep disorders to be relatively lower as 0.74±0.05 and 0.80±0.03 for LR and RF sleep models, respectively. Optimal sleep models trained using heterogeneous datasets perform very well when tested with the normal sleep dataset producing accuracy of ∼92%. Our study supports that using a more diverse training set benefits the sleep classifier model to be more generalizable for both healthy and abnormal sleepers.**

***Index Terms*— Modeling and analysis; Physiological monitoring - Modeling and analysis; Health monitoring applications**

I. Introduction

Accurate sleep assessment can provide useful information to indicate a subject's physical [1] and psychological health status [2]. Although polysomnography (PSG) is the gold standard of clinical sleep assessment, it is both complex, , and not suitable for short-term or long-term sleep monitoring. By contrast, modern wrist-worn actigraphy devices, capable of recording prolonged accelerometer data, are more simple and convenient for daily sleep monitoring [3], [4].

Although a number of actigraphy techniques have been reported to assess sleep quality in subjects with sleep disorders, several limitations preclude the adoption of such practices [5]. One of the major limitation is that, unlike typical study cohorts involving normal young sleepers, those involving a heterogeneous population that includes poor sleepers could showcase limited efficacy of actigraphy-based sleep assessment. Particularly, the estimated sleep epochs of abnormal sleepers could be more inaccurate for actigraphy algorithms involving simple rule based univariate approaches.

Authors are with Biofourmis Inc., Boston, MA 02110 USA. *Corresponding author, e-mail: kevin.chao@biofourmis.com

Several studies have successfully adopted machine learning algorithms on wrist-worn accelerometer data for sleep classification. One study applied four classifiers for sleep/wake detection among normal young sleepers [6]. Another recently published paper used a random forest classifier to detect the sleep patterns of a group of heterogeneous sleepers [4]. However, a systematic analysis on the composition and characteristics of study cohort influencing the machine learning models' efficacy for sleep assessment could be very valuable. Therefore, studying whether machine learning classifiers trained with a normal (or abnormal) sleeper dataset can successfully detect the sleep patterns of another abnormal (or normal) sleeper dataset requires further investigation.

The study presents actigraphy based multi-variate sleep detection models involving logistic regression (LR) and random forest (RF) classifiers, evaluates their performances systematically in both a healthy homogeneous population and a heterogeneous population with sleep disorders, and delineates the influence of study composition on the efficacy of sleep detection models using actigraphy.

II. Methods

A. Data

1) Newcastle Sleep Dataset [7]: This dataset contains left- and right-wrist tri-axial accelerometer data from 28 adult patients drawn from a one-night gold-standard polysomnography (PSG) experiment. The accelerometer data are recorded by a GENEActiv Watch at a sampling rate of 85.70 Hz. This study included only left-wrist accelerometer data from 27 subjects comprised of 8 normal sleepers and 19 abnormal sleepers. The average data length is 9.6±1.6 hours for PSG recordings and the average sleep length 69.0% over the entire dataset. The average age of the subjects is 45.6±14.4 years old.

2) Michigan 2019 Sleep Dataset [6], [8]: This dataset includes accelerometer and heart rate measurements by an Apple Watch and one night's sleep scored from the PSG recording. It contains 31 health subjects without any sleep disorder problems. The tri-axial accelerometer data are sampled at 50 Hz. The average data length for one subject is 7.2±1.5 hours for the PSG recording and the average sleep length 91.0% over the entire dataset. The average age of the subjects is 29.4±8.5 years old.

These public datasets were approved by the ethics committee and the institutional review board for the experimental procedures involved with the human subjects.

B. Actigraphy Features

Tri-axial accelerometer data were processed to extract the actigraphy features based on the feature type (see below for details), the calculation type (e.g., maximum, mean, sum, or standard deviation), and the input data resolution (1-, 15-, 30-, or 60-second time windows). The sample rates of 50 Hz and 85.7 Hz used in these two datasets are sufficient to capture all bodily motion [9], and the feature calculations are not impacted by the different sample rates. 61 features were computed for each subject, all of which were converted into 30-second epochs corresponding to 30-second sleep/wake PSG labels [6]. Figs. 1a-1e shows an example time series of motion and features for a health subject from the Newcastle dataset.

1) Activity Count [6], [10]: This feature refers to the actigraphy count obtained from the wrist watch. We computed the modified activity count in z-axis data by applying a 3-11 HZ band-pass filter and then divide the data into 128 bins between 0 and 5 g.

2) Activity Index (AI) [11]: The activity index feature, defined in Equation 1, is computed once per second from the variance of the three accelerometer axes:

$$AI = \sqrt{max\left(\frac{1}{3}\left\{\sum_{m=1}^{3}\frac{\sigma_{im}^2(t;H)-\bar{\sigma}_i^2}{\bar{\sigma}_i^2}\right\},0\right)} \qquad (1)$$

where $\bar{\sigma}_i^2 = \sigma_{ix}^2+\sigma_{iy}^2+\sigma_{iz}^2$ gives the noise-induced variance when the accelerometer is at rest. The 10th percentile of 10-second variances for each Michigan subject was computed, and mean of these values was assigned to $\bar{\sigma}_i^2$.

3) Signal Magnitude Area [12]: This feature is computed from the sum of the accelerometer magnitude over the tri-axial data and normalized by a given window length.

4) Inclination Angles [13]: These angles, which include X-Theta, Y-Psi, and Z-Phi, represent the orientation of the accelerometer device.

5) Z-angle [7]: We used the arm-angle change measurement of the approach as a feature to infer the active or inactive status of a subject.

6) Acceleration Magnitude [14]: After applying a 0.1 Hz highpass filter, each sampls's vector magnitude was calculated from the three accelerometer axes.

C. Sleep/Wake Classifier with Machine Learning

To create the sleep models, we applied feature elimination and tested the parameters to optimize each sleep classifier. Logistic regression (LR) and random forest (RF) models were tested for sleep/wake classification, defining a positive label (i.e., 1) as actual sleep and a negative label (i.e., 0) as wake.

First, backward elimination was run with all features using the ordinary-least-squares, retaining only features with p-values of less than 0.01. Next, we ran recursive feature elimination by maximizing the AUROC performance produced independent optimal feature sets for the intra-validation within the unique datasets (i.e., the Newcastle or Michigan) using LR and RF sleep models. The differences in feature

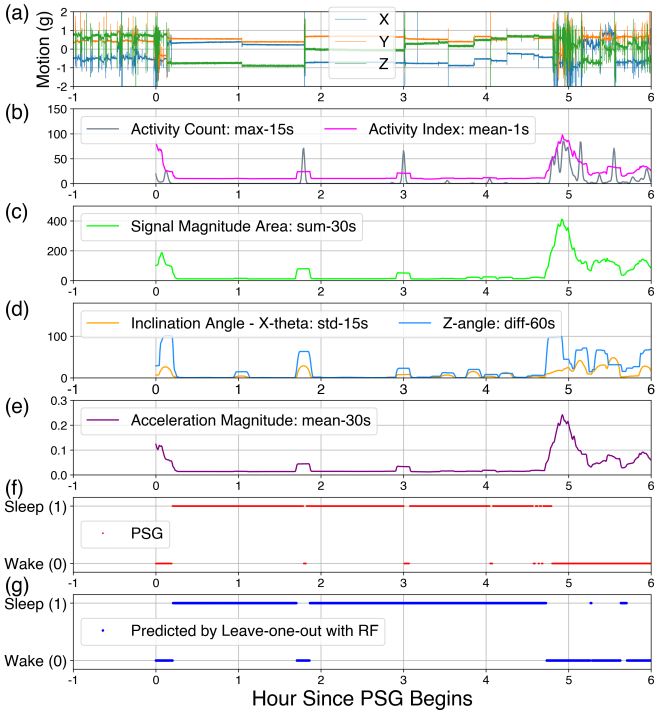

Fig. 1. Example of sensor data, actigraphy features, sleep annotation, and model performance (data from Newcastle subject mecsleep57-left, a normal sleeper). **(a)** Three-axis accelerometer data. **(b)-(e)** 6 representative actigraphy features. **(f)** PSG annotation. **(g)** Sleep and wake prediction from the RF model trained on all other Newcastle subjects.

sets are clearly influenced by the study cohort composition as well as the inherent differences between learning aspects of LR and RF models. However, the feature sets identified for the intra-validation of a given dataset and the ML model are retained the same for inter-validation analysis. This results in four distinct feature sets.

Finally, the selected features for each model were used to run 50 iterations of Monte Carlo cross-validation (70% training and 30% testing) and leave-one-out cross-validation [6].

For the metric to evaluate a model, we compared the PSG annotations with the sleep/wake epochs predicted by each model and then used the scikit-learn functions in Python to compute the median of AUC (area under the curve), accuracy, precision, sensitivity, F1 score, and specificity. Median AUC values calculated on the 50-run receiver operating characteristic curves were taken as the main metric for evaluating the performance of each model.

D. Cross-validation between the Newcastle and Michigan datasets

- Intra-dataset validation: First, 50 Monte Carlo train/test splits were run on each dataset (70% train and 30% test, split by subject). Next, leave-one-out cross-validation was performed, testing individually on each subject.
- Inter-dataset validation: Models were trained on the complete Newcastle dataset and tested individually on the subjects of Michigan. The process was then re-

versed, training on Michigan and testing on Newcastle by subject.

III. Results

A. Feature selection

The feature elimination procedure yielded two feature sets for the Michigan dataset (10 features for LR and 10 for RF) and two feature sets for the Newcastle dataset (9 features for LR and 24 for RF). Time series plots of six representative features are shown in Figs. 1b-1e. Fig. 1g presents an example of typical sleep/wake prediction from our models. The epochs classified as "wake" coincide with elevated feature amplitudes, indicating wristband motion.

B. Intra-dataset validation with leave-one-out and Monte Carlo

As shown in Table I, intra-dataset training and testing produce median values ranging from 0.736 (Monte Carlo validation on Newcastle LR model) to 0.893 (leave-one-out validation on Michigan RF model). The Michigan models show consistently higher median AUC values than the Newcastle models. Within the Michigan results, leave-one-out validation produces a ∼0.04 AUC improvement over Monte Carlo.

Within Michigan, Fig. 2 shows the individual ROC results for each validation run are more scattered for leave-one-out validation (Fig. 2, top row) than for Monte Carlo validation (Fig 2., bottom row). A comparison between the LR (left column) and RF models (right column) shows no significant difference between median AUC results and the distributions of the individual validation runs.

TABLE I

AUC value of intra-dataset validation

	Michigan		Newcastle	
	LR	RF	LR	RF
Leave-one-out	0.889±0.07	0.893±0.08	0.77±0.15	0.773±0.13
Monte Carlo	0.847±0.02	0.859±0.02	0.736±0.05	0.804±0.03

*median ± standard deviation

TABLE II

Performance by inter-dataset validation

	Test: Michigan (Train: Newcastle)		Test: Newcastle (Train: Michigan)	
	LR	RF	LR	RF
AUC	0.869±0.08	0.832±0.11	0.764±0.14	0.756±0.15
Accuracy	0.911±0.05	0.918±0.06	0.808±0.16	0.811±0.13
Precision	0.936±0.06	0.923±0.06	0.816±0.20	0.847±0.19
Sensitivity	0.961±0.04	1±0.01	0.988±0.08	0.906±0.17
F1 Score	0.952±0.03	0.957±0.03	0.892±0.16	0.874±0.18

*median ± standard deviation

The median AUC results of Newcastle with leave-one-out and Monte Carlo validation range between 0.736 and 0.804, notably less than the Michigan results. Leave-one-out validation (Fig. 3, top row) again shows little difference in median AUC performance and between the LR and RF models. The ROC results for individual normal and abnormal sleepers show similar distributions; the worst-performing ROCs, occasionally dropping below the diagonal, are typically associated with abnormally-sleeping subjects. In contrast, the bottom row of Fig. 3 shows Monte Carlo validation on Newcastle yields significantly better results from the RF model (median AUC 0.804) than from the LR model (median AUC 0.736).

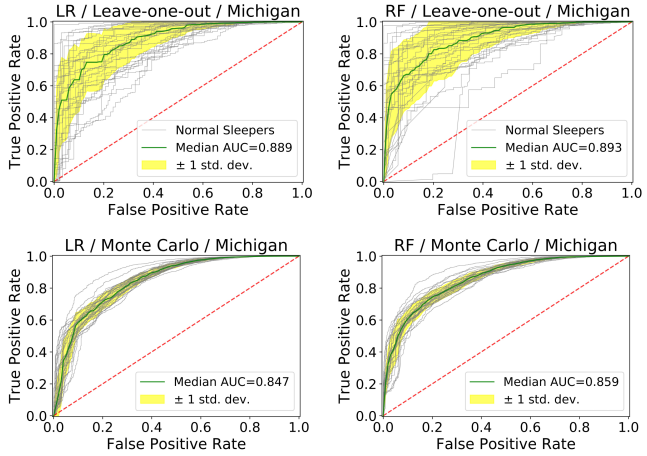

Fig. 2. Validation results for models trained and tested with the Michigan dataset: ROC curves for logistic regression sleep models (left column) and random forest sleep models (right column) evaluated using leave-one-out (top panels) and Monte Carlo (bottom panels) cross-validation methods. All subjects are normal sleepers.

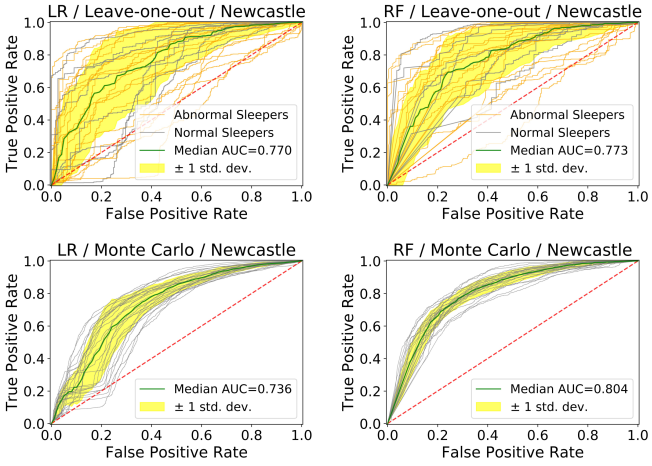

Fig. 3. Validation results for models trained and tested with the Newcastle dataset: ROC curves for logistic regression sleep models (left column) and random forest sleep models (right column) evaluated using leave-one-out (top panels) and Monte Carlo (bottom panels) cross-validation methods. The dataset consists of both normal and abnormal sleepers.

C. Inter-dataset validation

Fig. 4 presents the results for models trained with Newcastle and tested on individual Michigan subjects (top row), along with models trained on Michigan and tested with Newcastle subjects (bottom row). The median AUC results for the models tested on Michigan subjects (LR: 0.869

and RF: 0.832) are significantly better than those tested on Newcastle subjects (LR: 0.764 and RF: 0.756). The AUC results, along with the median subject values for other metrics, are summarized in Table II. The AUC results for the inter-dataset validation when tested on Michigan subjects (0.869 and 0.832) approximate those for intra-dataset validation on Michigan (Table I, all Michigan values $\geq$ 0.85). A similar trend is present for Newcastle data (Table I, where most Newcastle values are $\leq$ 0.8). The bottom row of Fig. 4 suggests relatively lower performance when tested on abnormal sleepers from Newcastle; a similar pattern is evident from intra-dataset validation on Newcastle (Fig. 3, top row).

As shown in Table II and Fig. 5, the accuracy, precision, and F1 are also generally better for inter-dataset validation with training with Newcastle and testing with Michigan. A comparison of the LR and RF models shows only modest differences in these metrics. The specificity (wake detection), however, shows significant differences between the LR and RF models. For inter-dataset testing on Michigan subjects, the LR model has stronger specificity (0.429±0.19) than does the RF model (0.24±0.2). The reverse is seen for testing on Newcastle, in which LR specificity (0.219±0.18) is weaker than then RF specificity (0.476±0.21).

IV. Discussion

The study presented multi-variate actigraphy-based logistic regression and random forest models for detecting sleep and investigating the influence of sleep characteristics in the performance of sleep detection. The results reveal stronger performances of AUC and F1 scores from intra- and inter-dataset cross validations in two unique sleep datasets than those of previous reports [4], [6].

A recent publication by Sundararajan et al. [4] examines a dataset consisting of 134 heterogeneous adult subjects (including the Newcastle subjects). This study trains an RF classifier with actigraphy-only features (i.e., the Z-angle, Euclidean Norm Minus One, and activity count). While the results of their study found an F1 score of 0.739 for sleep/wake classification, the F1 scores computed with RF classifiers in this study achieved higher F1 score values (0.874 and 0.957 in Table II).

While the current sleep models based on actigraphy features achieve high F1 scores for sleep detection, the specificity (i.e., correct detection of wake) is relatively lower (up to 0.476). This agrees with the literature [3] highlighting the inherent characteristics of actigraphy, overestimating sleep while underestimating awake states.

As far as the composition of study cohorts are concerned, the Newcastle dataset includes more diverse subjects with 70% of abnormal sleepers, while the Michigan dataset includes totally normal sleepers. This study composition significantly influences the proportion of awake data available for the training of sleep models. While the Michigan dataset contains 9% of awake time, Newcastle dataset is comprised of 32% of awake periods. Such imbalance in awake versus sleep periods resultant of the inherent differences in cohort

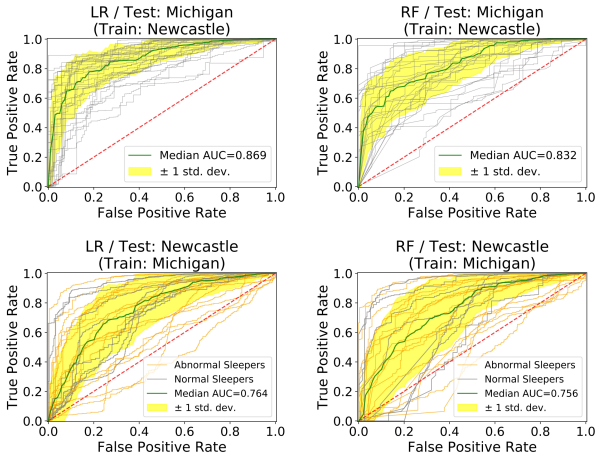

Fig. 4. Inter-dataset validation ROC curves for logistic regression (left column) and random forest (right column) models. Each ROC curve shows results for testing a single subject from one dataset (Newcastle or Michigan) using a model trained with all subjects from the other dataset.

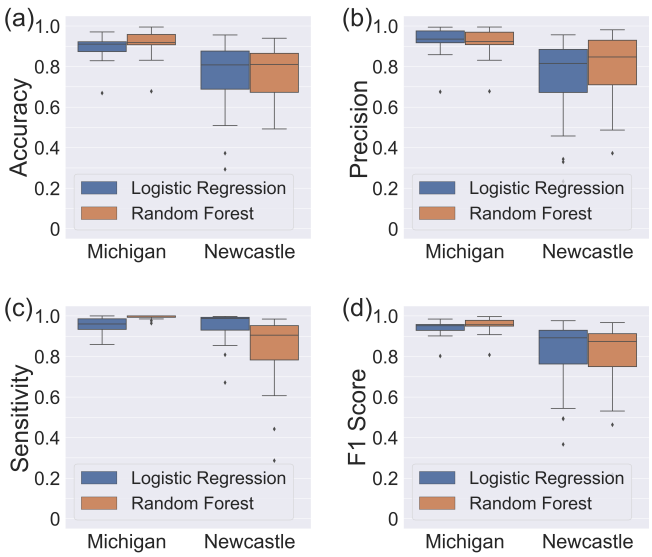

Fig. 5. Distributions of **(a)** accuracy, **(b)** precision, **(c)** sensitivity, and **(d)** F1 scores for the inter-dataset validation. Each distribution provides metrics for testing individual subjects from one dataset (Newcastle or Michigan) using a model trained with all subjects from the other dataset. The x-axis labels indicate the testing dataset. Median values are listed in Table II.

characteristic leads to showcasing substantial differences in both intra- and inter-dataset validation performance comparisons between the two datasets. At the same time, the choice of machine learning models between LR and RF does not produce noticeable differences in performances for a given dataset, as shown in Tables 1 and 2.

The performance of inter-dataset validation when training with Newcastle and testing with Michigan (Table II, AUC up to 87%) is roughly similar to intra-dataset validation on Michigan (Table I, AUC up to 89%). Conversely, inter-dataset validation when training with Michigan (Table II, AUC up to 76%) approximates the performance of intra-dataset validation for Newcastle (Table I, AUC up to 80%). In

both inter- and intra-dataset validation, we observed stronger performance when testing on the Michigan subjects. The awake time of the Newcastle subjects was longer than that of most of the Michigan subjects. This difference in performance of the models with the datasets could be attributed to the bias of our models while identifying sleep over wake time. Not surprisingly, these results also suggest that testing on a population with only healthy sleepers (Michigan) yields better results than using a test set that includes abnormal sleepers (Newcastle). Including abnormal sleepers in the training set, at least given the sample size available here, does not appear to completely overcome this trend. On the other hand, a training dataset including abnormal sleepers (Newcastle) may have more value for creating a comprehensive sleep model.

The future directions of research could further improve the sleep models by exploring additional actigraphy-based features, such as Euclidean Norm Minus One and Mean Amplitude Deviation [4], [15] that could enhance the diversity and predictive power of actigraphy features [16]. Given the merits of actigraphy such as inexpensiveness, convenience and long-term battery life, sleep monitoring using actigraphy remains a hot research topic. However, peripheral inactivity measured by actigraphy alone is known to be not adequate to differentiate sleep versus wakefulness, particularly in poor sleepers [17]. Therefore, enhanced actigraphy-based sleep features can complement to variety of physiological signals (e.g., PPG, respiratory, and heart rate variability) that can be simultaneously measured by the common wearable form factors including arm band, wrist watch, or wrist band. Such hybrid wearable sensing solutions and sleep models can overcome many of the present challenges for convenient and accurate sleep detection.

V. Conclusion

The study reveals that a diverse training dataset containing a heterogeneous population is crucial to produce better machine learning models for sleep detection than one containing only a homogeneous normal young population. Thus, the multi-variate sleep models trained with datasets comprised of heterogeneity could produce more robust and practical algorithms for sleep detection using actigraphy and other physiological features.

References

[1] K. A. Johnson et al., "The association of insomnia disorder characterised by objective short sleep duration with hypertension, diabetes and body mass index: A systematic review and meta-analysis," Sleep Medicine Reviews, vol. 59. 2021, doi: 10.1016/j.smrv.2021.101456.

[2] M. H. Hall, R. C. Brindle, and D. J. Buysse, "Sleep and cardiovascular disease: Emerging opportunities for psychology," Am. Psychol., vol. 73, no. 8, 2018, doi: 10.1037/amp0000362.

[3] M. De Zambotti, N. Cellini, A. Goldstone, I. M. Colrain, and F. C. Baker, "Wearable Sleep Technology in Clinical and Research Settings," Med. Sci. Sports Exerc., vol. 51, no. 7, 2019, doi: 10.1249/MSS.0000000000001947.

[4] K. Sundararajan et al., "Sleep classification from wrist-worn accelerometer data using random forests," Sci. Rep., vol. 11, no. 1, 2021, doi: 10.1038/s41598-020-79217-x.

[5] S. Ancoli-Israel, R. Cole, C. Alessi, M. Chambers, W. Moorcroft, and C. P. Pollak, "The role of actigraphy in the study of sleep and circadian rhythms," Sleep, vol. 26, no. 3. 2003, doi: 10.1093/sleep/26.3.342.

[6] O. Walch, Y. Huang, D. Forger, and C. Goldstein, "Sleep stage prediction with raw acceleration and photoplethysmography heart rate data derived from a consumer wearable device," Sleep, 2019, doi: 10.1093/sleep/zsz180.

[7] V. T. van Hees et al., "A novel, open access method to assess sleep duration using a wrist-worn accelerometer," PLoS One, 2015, doi: 10.1371/journal.pone.0142533.

[8] O. Walch, "Motion and heart rate from a wrist-worn wearable and labeled sleep from polysomnography (version 1.0.0)," PhysioNet, p. https://doi.org/10.13026/hmhs-py35, 2019.

[9] R. Khusainov, D. Azzi, I. E. Achumba, and S. D. Bersch, "Real-time human ambulation, activity, and physiological monitoring: Taxonomy of issues, techniques, applications, challenges and limitations," Sensors (Switzerland), vol. 13, no. 10. 2013, doi: 10.3390/s131012852.

[10] B. H. W. Te Lindert and E. J. W. Van Someren, "Sleep estimates using microelectromechanical systems (MEMS)," Sleep, 2013, doi: 10.5665/sleep.2648.

[11] J. Bai et al., "An Activity Index for Raw Accelerometry Data and Its Comparison with Other Activity Metrics," PLoS One, vol. 11, no. 8, p. e0160644, Aug. 2016, doi: 10.1371/journal.pone.0160644.

[12] D. M. Karantonis, M. R. Narayanan, M. Mathie, N. H. Lovell, and B. G. Celler, "Implementation of a real-time human movement classifier using a triaxial accelerometer for ambulatory monitoring," IEEE Trans. Inf. Technol. Biomed., 2006, doi: 10.1109/TITB.2005.856864.

[13] C. Fisher, "Using an accelerometer for inclination sensing," AN-1057, Appl. note, Analog Devices, 2010.

[14] D. M. Roberts, M. M. Schade, G. M. Mathew, D. Gartenberg, and O. M. Buxton, "Detecting sleep using heart rate and motion data from multisensor consumer-grade wearables, relative to wrist actigraphy and polysomnography," Sleep, 2020, doi: 10.1093/sleep/zsaa045.

[15] M. Marino et al., "Measuring sleep: Accuracy, sensitivity, and specificity of wrist actigraphy compared to polysomnography," Sleep, vol. 36, no. 11, 2013, doi: 10.5665/sleep.3142.

[16] S. Lüdtke, W. Hermann, T. Kirste, H. Beneš, and S. Teipel, "An algorithm for actigraphy-based sleep/wake scoring: Comparison with polysomnography," Clin. Neurophysiol., vol. 132, no. 1, 2021, doi: 10.1016/j.clinph.2020.10.019.

[17] W. W. Tryon, "Issues of validity in actigraphic sleep assessment," Sleep, vol. 27, no. 1. 2004, doi: 10.1093/sleep/27.1.158.