Publication:

Device agnostic sleep-wake segment classification from wrist-worn accelerometry
Luis R. Peraza, Richard Joules, Yves Dauvilliers, Robin Wolz
2020 IEEE International Conference on Healthcare Informatics (ICHI).

The published ICHI conference proceedings can be found in IEEExplore:
https://ieeexplore.ieee.org/xpl/conhome/1803080/all-proceedings

Device agnostic sleep-wake segment classification from wrist-worn accelerometry

Luis R. Peraza
IXICO Plc
London, UK
luis.peraza@ixico.com

Richard Joules
IXICO Plc
London, UK
richard.joules@ixico.com

Yves Dauvilliers
Department of Neurology
Centre Hospitalier Universitaire
Montpellier, France
y-dauvilliers@chu-montpellier.fr

Robin Wolz
[1]IXICO Plc
[2]Imperial College London
London, UK
robin.wolz@ixico.com

Abstract

Robustness for sleep-wake segment classification from accelerometry is critical when considering deployment in clinical studies. Deployed devices may change between or within studies, which alters critical clinical endpoints and negatively impact paired analyses if inter-device robustness is not assured. Here we present a neural network algorithm, deep learning sleep (DLS), for the classification of sleep-wake segments and show its robustness to different wrist-worn devices. Our results show that DLS delivers high accuracy when predicting sleep-wake segments in inter-device cross-validation experiments.

Index Terms

Deep learning, Parkinson's disease, wearables, clinical endpoint

I. Introduction

Sleep is disrupted by neurological diseases such as schizophrenia, Alzheimer's (AD), Parkinson's (PD), and Huntington's disease (HD) [1], [2]. Because of the close relation between sleep and disease, the accurate study of sleep would allow new ways for diagnosing diseases and assessing novel treatments. Highly accurate sleep studies are achieved by Polysomnography (PSG), a test in which the participant attends a sleep clinic for one or several nights for continuous sleep observation. However, PSG is expensive and can not be recorded for long periods (e.g. months). Alternative technologies such as wearable devices offer means to estimate an inexpensive proxy for PSG in long-term studies. Previous work on detecting sleep from accelerometry has developed algorithms to measure sleep time [3]–[5]. However, often these algorithms rely on hyperparameters that need to be tuned for a specific sensor device and disease. This device dependency undermines algorithm deployment in clinical trials where valid clinical endpoints must be obtained with high confidence.

In this paper, we present our deep learning sleep (DLS) algorithm for sleep detection from wrist-worn accelerometry and tested its performance on three databases that were recorded from different wearable devices and clinical populations. Our results show that DLS outperforms currently implemented sleep classification algorithms, Cole-Kripke (CK) [3] and Estimation of Stationary Segments (ESS) [5], especially in its ability to translate between devices.

II. Methods

Deep neural networks have shown great potential for sleep classification from accelerometry as well as better performance than feature-designed algorithms [6]–[8]. Our proposed DLS method was developed with Tensorflow-Keras (v2.2.4) and is fed with z-axis accelerometry segments through two input branches: one for band-pass filtered accelerometry and the second for FFT coefficients. The accelerometry branch comprises five CNN-1D layers and the FFT branch two dense layers. Both branches are subsequently concatenated and become an input of three dense layers with a final perceptron as output.

A. Accelerometry Databases

We analysed three databases with ground-truth PSG: The CONTEXT study (IXI) [6], the Technische Universität Darmstadt (TUD) [5], and the Newcastle PSG (NCL) [9] databases. The IXI database comprises accelerometry acquired with AX3 devices (axivity.com) from 46 participants, ten of them diagnosed with PD [6].

The TUD database is described in [5] and comprises accelerometry recordings (HedgeHog wrist-worn device) from participants who were diagnosed with varied sleep disorders: sleep apnea syndrome (SAS), restless leg syndrome (RLS), insomnia, narcolepsy, and REM behavioural disorder (RBD). The NCL database is a similar database and was recorded with the GENEActiv device (www.activinsights.com), see Table I.

TABLE I
ACTIGRAPHY DATABASE SUMMARY.

variable	**IXI**	**TUD**	**NCL**
Wearable device	AX3	HedgeHog	GENEActiv
No. of recordings	46	46	53
Healthy controls	36	5	8
Patients	10	38	20
Mean age (SD)	87.78 (4.28)	58.32 (15.04)	44.86 (14.65)
Data (in hours)	370.18	396.66	498.13

B. Experimental design and model training

The algorithms were assessed using a cross-validation design with independent evaluation tests. For this, two databases were chosen for cross-validations, i.e. training with database A and testing on database B, and vice versa. The third database, database C, was used for independent evaluation. We chose the IXI and TUD databases for cross-validation and the NCL database for independent evaluation. Cross-validation resulted in two trained models per algorithm; e.g. for the DLS algorithm these models would be IXI-DLS and TUD-DLS, which were respectively cross-validated and labelled as IXI-DLS-TUD and TUD-DLS-IXI. The independent evaluations using the NCL database were labelled as IXI-DLS-NCL and TUD-DLS-NCL. This labelling scheme was also followed for the other two algorithms; CK and ESS.

For the DLS training, the database was divided in train (99%) and test (1%) sets across all accelerometry segments; batch size was 50, and earlystopping was used to define convergence. The CK and ESS algorithms were tuned using a grid search for the ideal hyperparameters. After grid-searching, the hyperparameters that resulted in the highest accuracy for the training database were stored and used in the cross-validation and independent evaluation experiments.

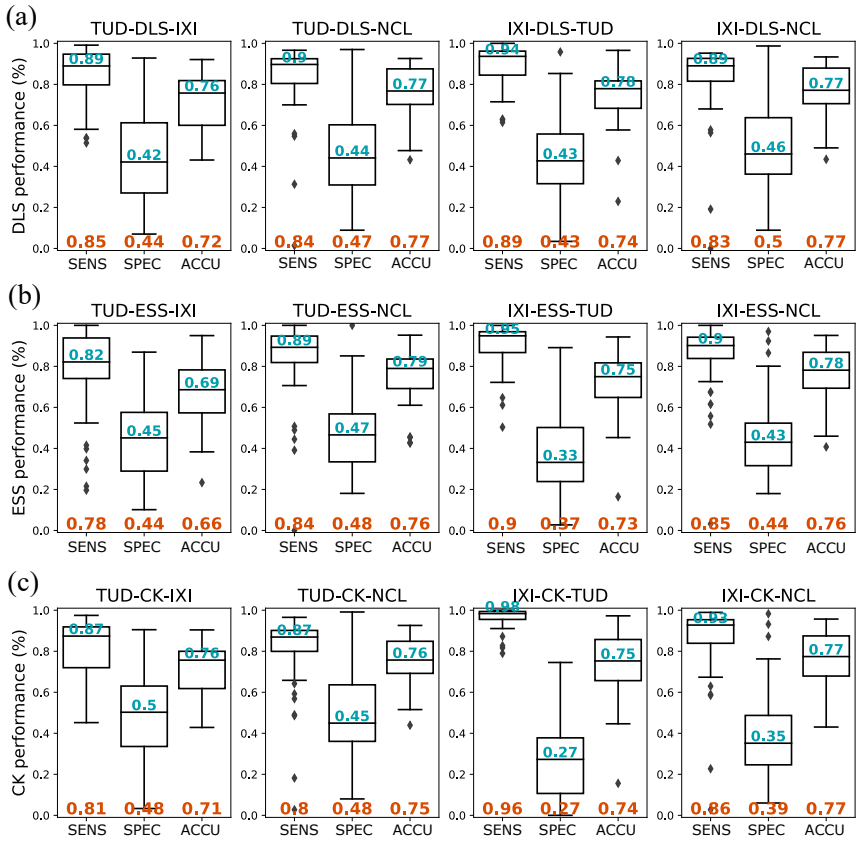

Fig. 1. Sleep-wake segment classification; cross-validations and independent evaluations. (a) Cross-validation and evaluation results for the DLS algorithm, (b) and (c) same as (a) but for the ESS and CK algorithms. Each box plot shows the performance median value within the boxes and the mean at its bottom. Sensitivity (SENS), specificity (SPEC), and accuracy (ACCU).

III. Results

Cross-validation and independent evaluation results are shown in Fig. 1. The mean accuracy of the DLS was higher than ESS and CK algorithms in all cross-validation and independent evaluations. This difference was higher when comparing the mean accuracy between the TUD-DLS-IXI and TUD-ESS-IXI cases, with mean accuracies of 0.72 and 0.66 respectively. Additionally, DLS showed a more stable performance across cross-validations and evaluations, with high sensitivity ($>= 0.89$) and specificity ($>= 0.42$). On the contrary, ESS reached a specificity of 0.33 for the IXI-ESS-TUD case and the CK algorithm reached a mean specificity of 0.27 for the equivalent case.

Fig. 2-top shows sleep-wake segment estimation results from a healthy participant within the NCL database. Predictions were made using IXI-trained models: IXI-DLS, IXI-CK and IXI-ESS. Overall the three models estimated sleep and wake segments correctly, however the CK algorithm failed in estimating a wake segment around 03:48 hours.

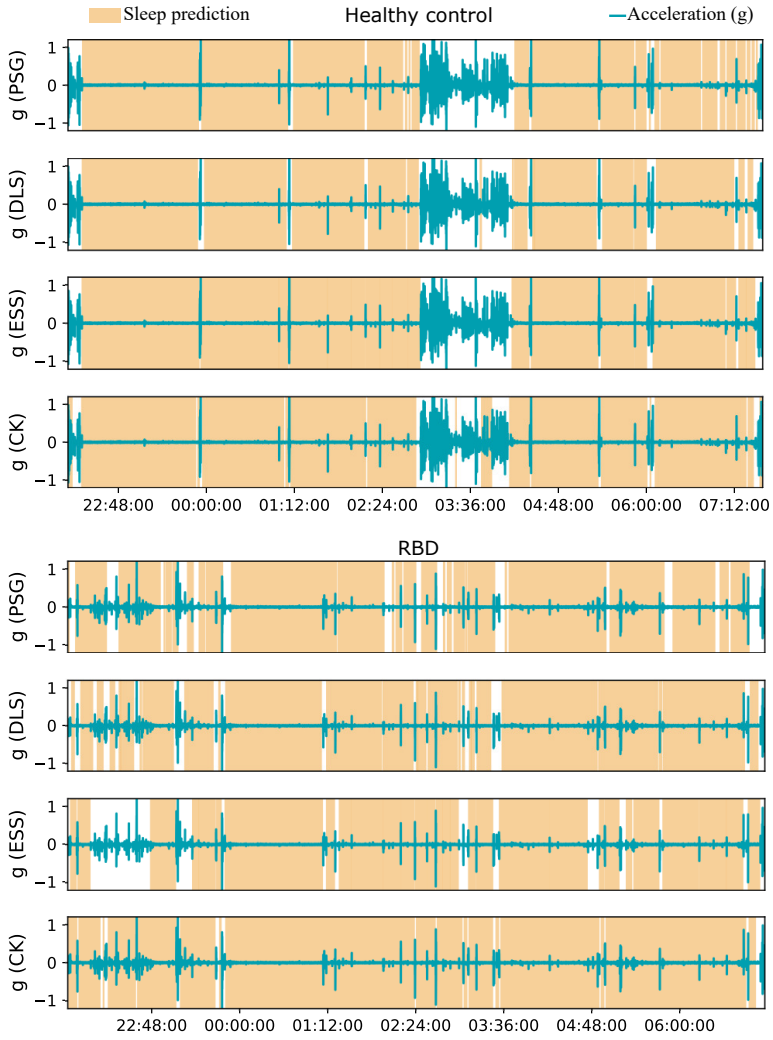

Fig. 2. Sleep-wake segment estimation by the three algorithms for a healthy control and RBD participants from the Newcastle (NCL) database. Ground-truth polysomnography (PSG) is shown at the top of each panel.

A classification example from an RBD patient is shown in Fig. 2-bottom. RBD is a sleep disorder where an individual acts out their dreams; patients may move their limbs, talk and even walk out of the bed. This makes sleep estimation challenging, however, DLS and CK showed a good agreement with PSG. ESS on the other hand misclassified an important time segment as awake.

IV. Conclusion

Our results showed that DLS outperformed two widely used algorithms for sleep segment classification, ESS and CK. The DLS algorithm showed higher mean accuracy in all experiments. Noticeable from our experiments is that the mean values for sensitivity, specificity and accuracy for the DLS did not significantly change across cross-validation and independent evaluation

experiments, i.e. DLS performance did not change across wearable devices. Device agnosticism and robustness across different patient populations is crucial when high confidence in the deployed algorithms is necessary, as is the case in real-world evidence assessment for clinical trials.

REFERENCES

[1] M. Van Egroo, J. Narbutas, D. Chylinski, P. Villar Gonzalez, P. Maquet, E. Salmon, C. Bastin, F. Collette, and G. Vandewalle, "Sleep-wake regulation and the hallmarks of the pathogenesis of Alzheimer's disease," Sleep, vol. 42, April 2019.

[2] E. Sinforiani, R. Zangaglia, R. Manni, S. Cristina, E. Marchioni, G. Nappi, F. Mancini, and C. Pacchetti, "REM sleep behavior disorder, hallucinations, and cognitive impairment in Parkinson's disease," Mov Dis, vol. 21, pp. 462-466, 2006.

[3] RJ. Cole, DF. Kripke, W. Gruen, DJ. Mullaney, and C. Gillin, "Automatic sleep/wake identification from wrist activity," Sleep, vol. 15, pp. 461-469, 1992.

[4] M. Takeshima, M. Echizenya, Y. Inomata, K. Shimizu, and T. Shimizu. "Comparison of sleep estimation using wrist actigraphy and waist actigraphy in healhty young adults," Sleep Biol Rhythms, vol. 12, pp. 62-68, 2014.

[5] M. Borazio, E. Berlin, N. Kücükyildiz, P. Scholl, and K. Van Laerhoven, "Towards benchmarked sleep detection with inertial wrist-worn sensing units," IEEE International Conference on Healthcare Informatics. Verona, 2014, pp. 125-134.

[6] R. Wolz, J. Munro, R. Guerrero, DL. Hill, and Y. Dauvilliers, "Predicting sleep/wake patterns from 3-axis accelerometry using deep learning," Alzheimer's & Dementia, vol. 13, P1012, 2017.

[7] R. Wolz, J. Munro, R. Guerrero, D. Hill, and Y. Dauvilliers, "Extracting digital biomarkers of sleep from 3-axis accelerometry using deep learning," J Prev Alzheimer's Dis, vol. 4, P81, 2017.

[8] K. Kinnunen, R. Joules, J. Munro, I. Simpson, R. Wolz, and Y. Dauvilliers, "Comparison of sleep measurements from actigraphy to self-reported sleep diaries," J Prev Alzheimer's Dis, vol. 5, P146, 2018.

[9] V. van Hees, S. Sabia, SE. Jones, AR. Wood, KN. Anderson, M. Kivimäki, TM. Frayling, AI. Pack, M. Bucan, MI. Trenell, DR. Mazzotti, PR, Gehrman, BA. Singh-Manoux, and MN. Weedon, "Estimating sleep parameters using an accelerometry without sleep diary," Sci Rep, vol. 8, 12975, 2018.

[10] F. Chollet and others, "Keras: The python deep learning library", https://keras.io, 2015.