Reliability of wrist-worn accelerometry devices and algorithms for sleep detection

Luis R Peraza[1], Richard Joules[1], Yves Dauvilliers[2], Robin Wolz[1]
1) IXICO Plc – London UK, 2) Centre Hospitalier Universitaire – Montpellier France

Introduction

Wearable devices have shown great potential for obtaining measures of real-world evidence in clinical trials, but standardization and variability between different devices remains one of the barriers for systematic deployment. In this investigation, we present a neural network algorithm for sleep detection, Deep Learning Sleep (DLS), and compare its reliability across different datasets and two widely used sleep algorithms: Cole-Kripke (CK) [1] and the estimation of stationary segments algorithm (ESS) [2]. Our results show that DLS algorithm outperforms CK and ESS by delivering higher mean accuracy when predicting sleep-wake segments in inter-device cross-validation experiments.

Methodology

To test the performance of the three sleep algorithms, we used three accelerometry databases; one internal and two publicly available: The CONTEXT study database (IXI, ixico.com), The Technische Universität Darmstadt database (TUD) [2] and the Newcastle polysomnography database (NCL) [3]. The IXI database comprises accelerometry data acquired with AX3 devices (axivity.com) from 46 participants, ten of them diagnosed with Parkinson's disease (PD), who were recruited within a community study at the Centre Hospitalier Universitaire in Montpellier France, see Table I.

TABLE I
ACTIGRAPHY DATABASES.

variable	IXI	TUD	NCL
Wearable device	Axivity	HedgeHog	GENEActiv
No. of recordings	46	46	53
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

For the DLS training, the training 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.

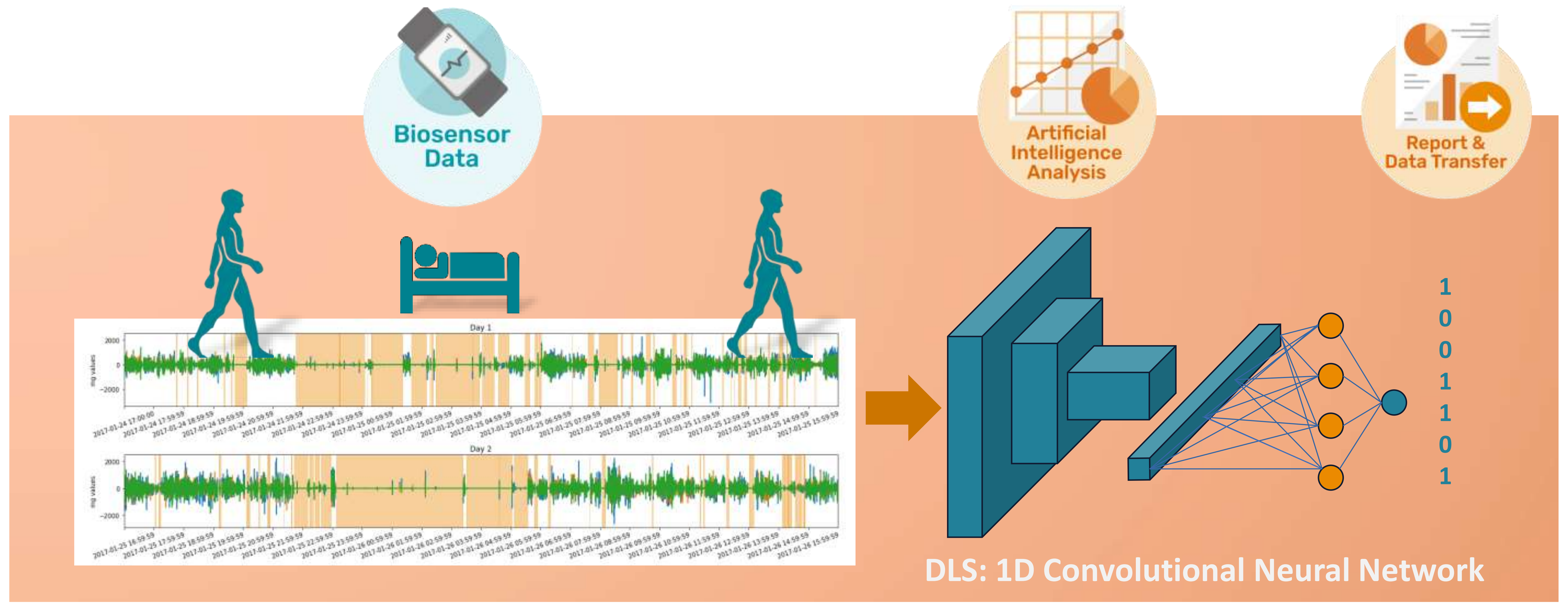

Figure 1. Data analysis pipeline. Recorded datasets are pre-processed and fed to the DLS algorithm for feature extraction and sleep-wake output predictions.

Results

Cross-validation and independent evaluation results are shown in Fig. 2. 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 [4].

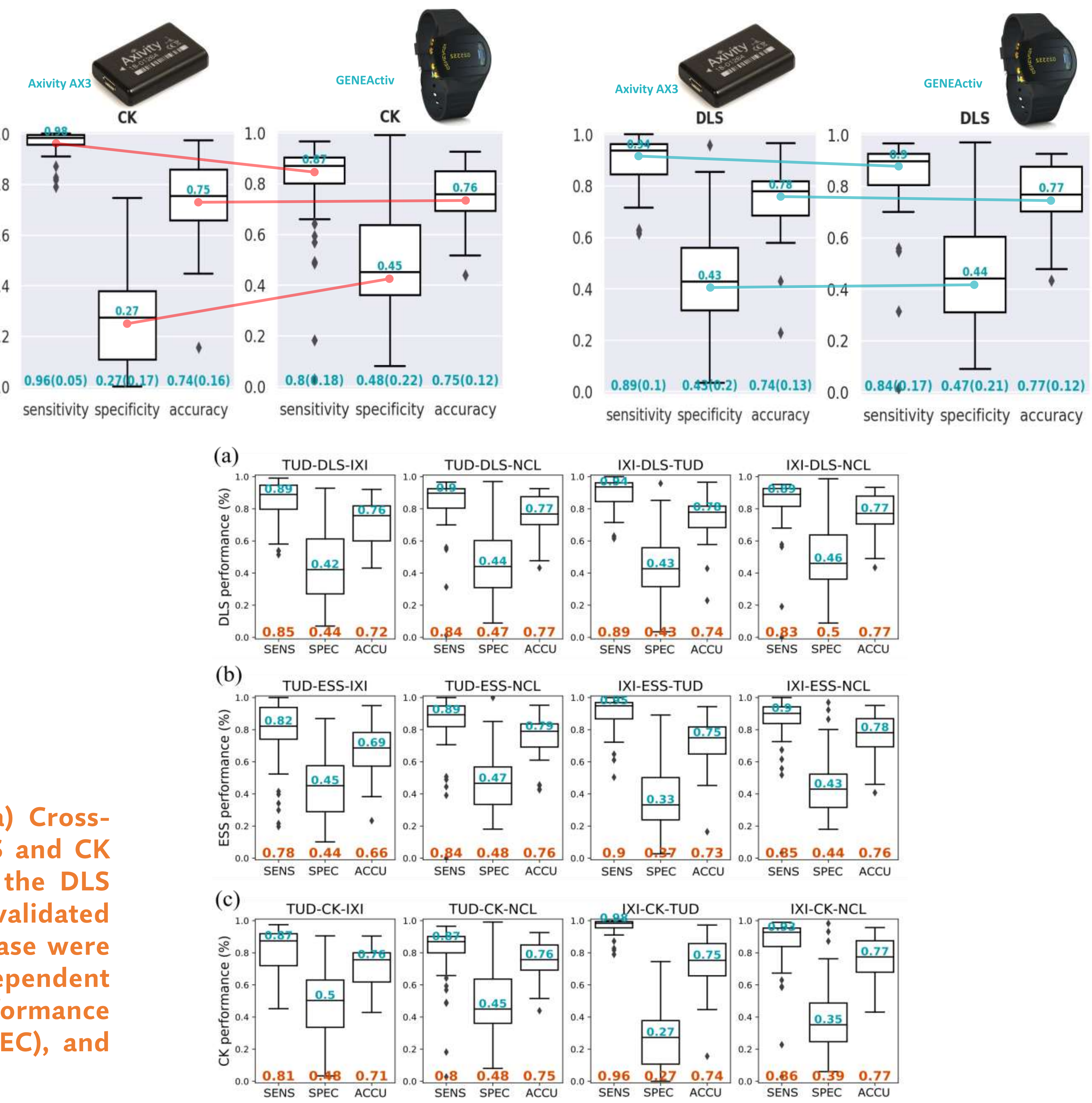

Figure 2. Sleep-wake segment classification; cross-validations and independent evaluations. (a) Cross-validation and evaluation results for the DLS algorithm, (b) and (c) same as (a) but for the ESS and CK algorithms. Cross-validation resulted in two trained models per algorithm; for example, for the DLS algorithm these trained 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 with the NCL database were labelled as IXI-DLS-NCL and TUD-DLS-NCL. This labelling scheme for cross-validation and independent evaluation was followed for the other two algorithms, CK and ESS. Each box plot shows the performance median value within the boxes and the mean at its bottom. Sensitivity (SENS), specificity (SPEC), and accuracy (ACCU).

Fig. 3-left 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.

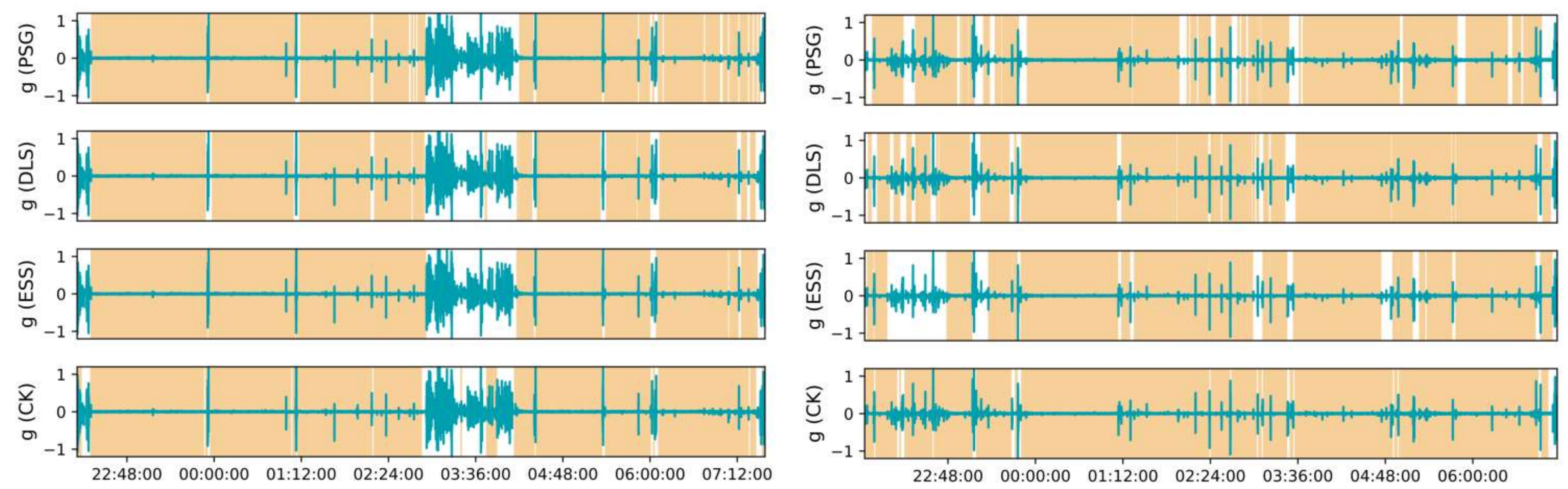

Figure 3. Sleep-wake segment estimation by the three algorithms for a healthy control and RBD participants from the Newcastle (NCL) database. Ground-truth polysomnography (PSG) is shown at the top of each panel.

A classification example from an RBD (REM behavioural disorder) patient is shown in Fig. 3-right. 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.

Conclusions

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] Cole RJ, Kripke DF. "Automatic sleep/wake identification from wrist activity," Sleep, vol. 15, pp. 461-469, 1992.

[2] Borazio M, Berlin E, et al. "Towards benchmarked sleep detection with inertial wrist-worn sensing units," IEEE ICHI 2014.

[3] van Hees V, Sabia S, et al. "Estimating sleep parameters using an accelerometer without sleep diary," Sci Rep, vol. 8, 12975, 2018.

[4] Peraza LR, Joules R, Dauvilliers Y, Wolz R. "Device agnostic sleep-wake segment classification from wrist-worn accelerometry," IEEE ICHI 2020 (accepted).