Assessing the Performance of Sleep Staging Models Trained on Simulated Data Using Wearable Behind-the-Ear EEG Signals

Gert Vanhollebeke [1], Nigel Colenbier [1], Emiel Vereycken [1], Caroline Neuray [2], Hans Danneels [3], Riem El Tahry [4,5], Pieter van Mierlo [1]

[1] Clouds of Care NV, Ghent, Belgium
[2] Center for Cognitive Neuroscience, Salzburg, Austria
[3] Byteflies NV, Antwerp, Belgium
[4] Institute of Neuroscience, Université Catholique de Louvain (UCLouvain), Brussels, Belgium
[5] Centre for Refractory Epilepsy, Cliniques Universitaires Saint-Luc, Brussels, Belgium

clouds of care

Byteflies

INTRODUCTION

- Children with epilepsy often have co-occurring sleep disorders yet sleep architecture properties are an under-evaluated aspect of the disease profile in this population [1], [2].
- Polysomnography is commonly used for sleep analysis, but the multitude of electrodes used in the full PSG setup are sometimes not well tolerated in pediatric epilepsy patients.
- Wearable behind-the-ear (BTE) EEG devices hold significant potential for analyzing sleep in pediatric epilepsy patients as their minimal electrode setup is less intrusive. Additionally, BTE EEG devices can be used to record sleep data at home, making multi-night sleep analysis more accessible .
- Data from BTE EEG devices, however, are difficult to score visually and thus highlights the need for automated sleep staging. Labeled datasets from these devices, however, are currently lacking which makes the construction of automated sleep staging models difficult.
- One potential solution for this problem is to train models on simulated BTE data, based on near-ear scalp EEG electrodes, and evaluate the performance on real wearable BTE device data (see Figure 1 for a visual representation of the proposed approach).
- In this study, we explore the feasibility of this approach by comparing model performance between simulated and wearable BTE data in a pediatric epileptic dataset.

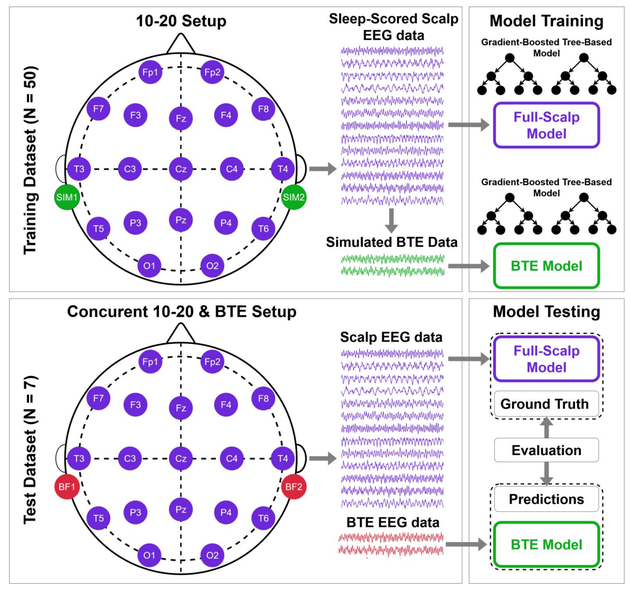

Figure 1: Overview of proposed modeling approach. *Upper part*: usage of the training dataset for BTE simulation and model training. *Lower part*: usage of the test set for ground truth generation and BTE model evaluation.

METHODS

- A dataset of 50 24-hour EEG recordings with 19 electrodes (10-20 system) of pediatric patients without epileptic abnormalities during the recording (mean age = 7.85 ± 4.07 years) from Saint-Luc University Hospital (Brussels, Belgium) was used for model training.
- Each recording was scored by three experts, and consensus scores (i.e., when at least two of the three experts agreed) were used as labels to train models to classify Wake, NREM1, NREM2, NREM3 and REM sleep stages.
- Two models were trained, one on full-scalp EEG (Full-scalp model) and one on a simulated BTE signal (BTE model; formula = (T3 – T4) * 0.3; obtained through ordinary least-squares estimation (see Figure 2)).
- Both models were gradient boosting tree-based models, which were trained on features obtained from 30 second epochs of either full-scalp or simulated BTE EEG data. Used features were spectral power, complexity, and statistical property features.
- For model evaluation, a test dataset of 7 overnight recordings with simultaneously recorded full-scalp- and wearable device-recorded data was used. The wearable device collected both BTE and T3-T4 electrode data.
- Sleep stage predictions provided by the full-scalp model were used as ground truth for the test dataset. Three predictions of the BTE model were evaluated: predictions based on simulated BTE data, predictions based on the wearable device-recorded T3-T4 electrode data, and predictions based on the wearable device-recorded BTE data.
- For each signal, predictions on three levels of detail regarding sleep architecture were evaluated: minimum (Wake vs. Sleep), medium (Wake vs. NREM vs. REM), and maximum (Wake vs. NREM1 vs. NREM2 vs. NREM3 vs. REM). Model agreement was assessed using Cohen's Kappa (κ).

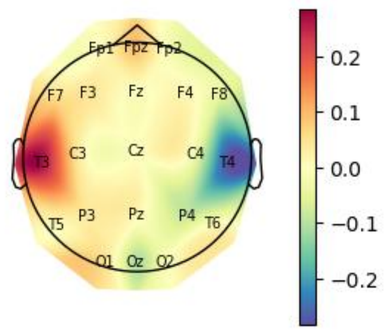

Figure 2: Average OLS solution for simulation of BTE montage from the full-scalp EEG setup.

RESULTS

- For the training set, the interrater agreement is substantial (κ = 0.87 ± 0.06), as is the agreement of the full-scalp model with the expert ratings (κ = 0.87 ± 0.09).
- Predictions based on the simulated BTE signal have substantial agreement on all three resolution levels (Figure 3) (*minimum*: κ = 0.67 ± 0.19; *medium*: κ = 0.7 ± 0.11; *maximum*: κ = 0.67 ± 0.13)
- Prediction based on the wearable T3-T4 signal have moderate agreement on all three resolution levels (Figure 3) (*minimum*: κ = 0.54 ± 0.25; *medium*: κ = 0.55 ± 0.15; *maximum*: κ = 0.5 ± 0.16)
- Predictions based on the wearable BTE signal have moderate agreement on the minimum and medium resolution, and fair agreement on the maximum level (Figure 3) (*minimum*: κ = 0.41 ± 0.22; *medium*: κ = 0.42 ± 0.21; *maximum*: κ = 0.34 ± 0.21)

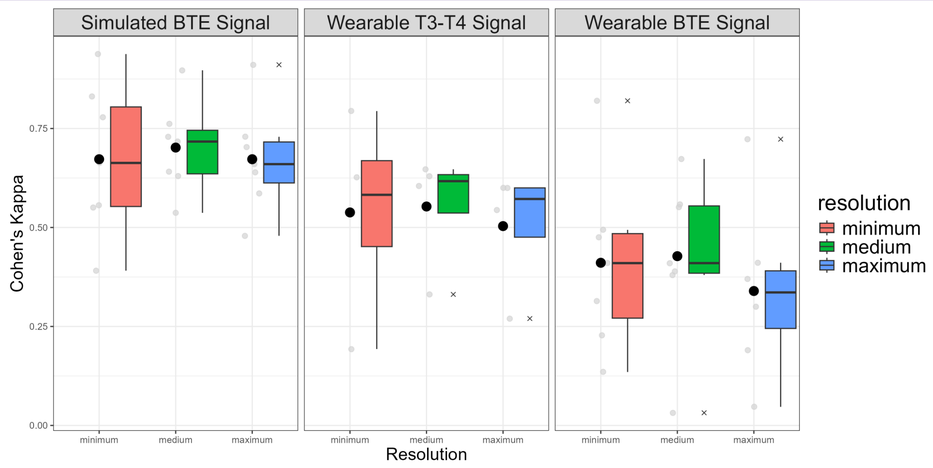

Figure 3: Results across evaluated signals and resolutions. The large black dots indicate the mean Cohen's Kappa.

CONCLUSION

- Predictions based on the simulated BTE signal in the test dataset have substantial agreement with the ground truth, indicating that the BTE model can identify sleep stages from electrodes not used in the standard PSG for sleep staging (i.e., frontal, central, and occipital electrodes).
- Across the evaluated signals, it is noticeable that minimum and medium resolution have similar performance, while the maximum resolution has slightly worse performance. This hints at the fact that the BTE model can identify Wake, NREM and REM sleep stages, but that the different NREM stages are more difficult to separate. The likely reason for this is that some NREM stages are defined by EEG features most pronounced in frontal and central electrodes (i.e., K-complexes and spindles) that are less represented in temporal or BTE electrodes.
- Predictions using the signal that was used for model training (i.e., the simulated BTE signal) result in the best performance in the test set while predictions from the wearable BTE signal in the worst performance, showing that differences between the simulated and real BTE signal affect model performance.
- These results indicate the potential of using BTE simulations to train sleep staging models for wearable BTE EEG devices but also highlight the fact that deviating from the simulated data used for model training results in worse performing models.

References

[1] Stores, Wiggs, & Campling. (1998). Sleep disorders and their relationship to psychological disturbance in children with epilepsy. *Child: care, health and development*, *24*(1), 5-19.

[2] Manni, R., & Terzaghi, M. (2010). Comorbidity between epilepsy and sleep disorders. *Epilepsy research*, *90*(3), 171-177.