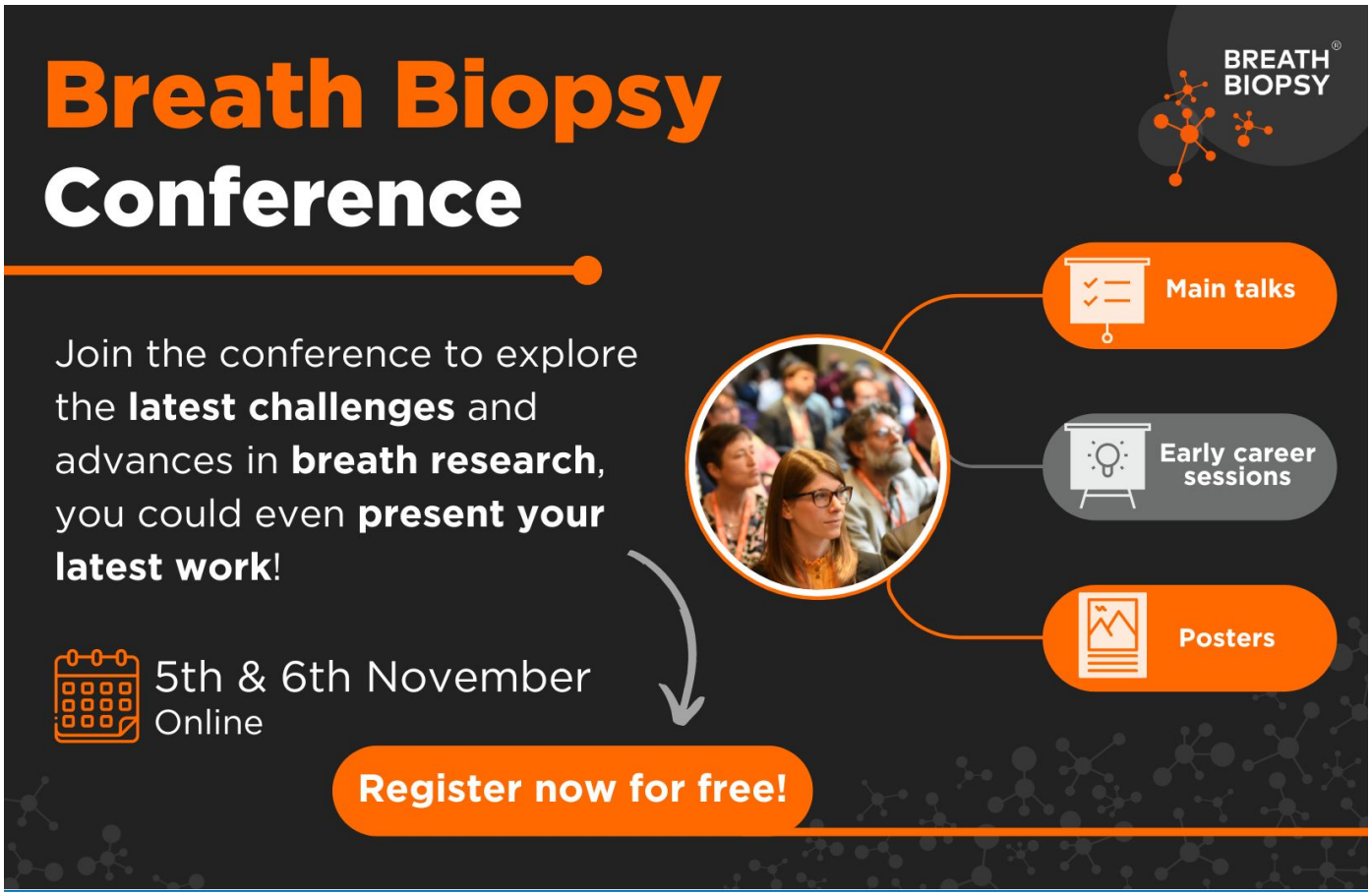
PAPER • **OPEN ACCESS**

SleepSEEG: automatic sleep scoring using intracranial EEG recordings only

To cite this article: Nicolás von Ellenrieder *et al* 2022 *J. Neural Eng.* **19** 026057



Journal of Neural Engineering

PAPER

SleepSEEG: automatic sleep scoring using intracranial EEG recordings only

Nicolás von Ellenrieder[1,*], **Laure Peter-Derex**[1,2,3], **Jean Gotman**[1] **and Birgit Frauscher**[1]

[1] Montreal Neurological Institute and Hospital, McGill University, Montreal, Canada
[2] Center for Sleep Medicine and Respiratory Diseases, Hospices Civils de Lyon, Lyon 1 University, Lyon F-69000, France
[3] Lyon Neuroscience Research Center; CNRS, UMR5292; INSERM, U1028, Lyon F-69000, France
[*] Author to whom any correspondence should be addressed.

E-mail: nicolas.vonellenrieder@mcgill.ca

Keywords: focal epilepsy, sleep staging, stereo-electroencephalography, local sleep, automatic classification

Supplementary material for this article is available online

OPEN ACCESS

RECEIVED
22 December 2021

REVISED
6 April 2022

ACCEPTED FOR PUBLICATION
18 April 2022

PUBLISHED
3 May 2022

Abstract

Objective. To perform automatic sleep scoring based only on intracranial electroencephalography (iEEG), without the need for scalp EEG), electrooculography (EOG) and electromyography (EMG), in order to study sleep, epilepsy, and their interaction. *Approach.* Data from 33 adult patients was used for development and training of the automatic scoring algorithm using both oscillatory and non-oscillatory spectral features. The first step consisted in unsupervised clustering of channels based on feature variability. For each cluster the classification was done in two steps, a multiclass tree followed by binary classification trees to distinguish the more challenging stage N1. The test data consisted in 11 patients, in whom the classification was done independently for each channel and then combined to get a single stage per epoch. *Main results.* An overall agreement of 78% was observed in the test set between the sleep scoring of the algorithm using iEEG alone and two human experts scoring based on scalp EEG, EOG and EMG. Balanced sensitivity and specificity were obtained for the different sleep stages. The performance was excellent for stages W, N2, and N3, and good for stage R, but with high variability across patients. The performance for the challenging stage N1 was poor, but at a similar level as for published algorithms based on scalp EEG. High confidence epochs in different stages (other than N1) can be identified with median per patient specificity >80%. *Significance.* The automatic algorithm can perform sleep scoring of long-term recordings of patients with intracranial electrodes undergoing presurgical evaluation in the absence of scalp EEG, EOG and EMG, which are normally required to define sleep stages but are difficult to use in the context of intracerebral studies. It also constitutes a valuable tool to generate hypotheses regarding local aspects of sleep, and will be significant for sleep evaluation in clinical epileptology and neuroscience research.

1. Introduction

Intracranial electroencephalography (iEEG), performed in drug-resistant focal epilepsy patients undergoing presurgical evaluation, represents a unique opportunity to study sleep from direct cortical recordings (Frauscher and Gotman 2019). They provide insights into the mechanisms of epilepsy and sleep, including their multiple and reciprocal interactions (Frauscher and Timofeev 2021). Both ictal and interictal activity are modulated by sleep (Ng and Pavlova 2013, Frauscher *et al* 2015, 2016, Campana *et al* 2017), and at the same time, both ictal and interictal activity can impact sleep structure and function (Terzaghi *et al* 2008, Halasz *et al* 2019, Lambert *et al* 2020, Peter-Derex *et al* 2020). Moreover, such recordings are of great interest as they allow to explore local aspects of sleep and to provide valuable information about sleep physiology (Magnin *et al* 2004, 2010, Nir *et al* 2011, Peter-Derex *et al* 2012, 2015, Frauscher *et al* 2020).

Sleep scoring is a way to classify the vigilance state of the brain in different stages defined based on scalp EEG, electrooculogram (EOG), and chin

electromyogram (EMG) signals. It is standardized by sleep scoring guidelines developed by the American Academy of Sleep Medicine (Berry *et al* 2017), and used for studying sleep. While pediatric sleep scoring is similar to adults, sleep scoring in neonates differs greatly (Dereymaeker *et al* 2017). In adult subjects, epochs of 30 s duration are defined through the night, and one of five possible sleep stages are assigned to each epoch based on the characteristics of the scalp EEG, ocular activity, and chin muscle tone. These stages are W for wakefulness, R for rapid-eye-movement (REM) sleep, and N1 to N3 for non-REM sleep, ranging from N1 representing drowsiness to N3 corresponding to deep sleep.

iEEG is performed as part of presurgical investigation of drug-refractory focal epilepsy patients. Depth electrodes are inserted in the brain or grids and strips placed on its surface. While for scalp EEG the low electric conductivity of the brain leads to a spatial blurring of the cortical activity (Hämäläinen *et al* 1993), the iEEG measurements are much more local, recording from less than a centimeter of the electrodes (von Ellenrieder *et al* 2021). Also, with iEEG the activity of deep structures such as the insula or mesio-temporal brain regions can be recorded, whereas this activity is usually not visible in the scalp EEG. Therefore, scalp EEG and iEEG can be considered as different modalities even if both measure the electric activity of the brain. In fact, clinicians used to reading scalp EEG require specific training to interpret iEEG studies.

Sleep studies using iEEG face several challenges. Presurgical evaluations typically last 1 week or longer, generating a large amount of data. To study sleep based on this data, sleep scoring is essential, but it is challenging and time consuming when performed manually. When scalp EEG is recorded simultaneously, the montage is rarely standard given the variable extent and location of the implantation, which is tailored to a personalized clinical hypothesis. Furthermore, collodion glued standard scalp electrodes, EOG, and EMG lead not only to increased patient discomfort, but also inconsistent signal quality, since contacts may be lost during the night and not recovered, as the patient is not in a controlled sleep lab environment. Moreover, to obtain a reasonable data quality, re-gluing of electrodes requires removal of the bandage protecting the depth electrodes, further complicating the clinical set-up. Another drawback of using a conventional scalp EEG sleep montage is the reference to the mastoid following standard criteria of the American Academy of Sleep Medicine (Berry *et al* 2017). A mastoid reference is often a poor reference in case of temporal lobe epilepsy where contamination of this reference would naturally occur because of interictal epileptiform discharges (IEDs) (Marzec and Malow 2003). Finally, manual scoring of the numerous recorded nights is time-consuming given the mentioned technical challenges and pathological brain activity, resulting in a limited number of patients usually included in iEEG sleep studies. Thus, an automatic scoring algorithm based on iEEG alone would be of great practical interest to facilitate the clinical workflow, process greater amount of data, and guarantee reproducibility of the scoring results, as required especially in multicenter studies. Ideally, such an algorithm would also be a tool to aid in the study of local sleep characteristics of different cortical regions, as emerging evidence underlines that various brain regions and in particular deep structures sleep differently from the rest of the brain (Sarasso *et al* 2014, von Ellenrieder *et al* 2020, Olejarczyk *et al* 2022).

Due to the mentioned challenges as well as the unmet clinical need, we developed SleepSEEG, an automatic sleep scoring algorithm based on iEEG, with minimal requirements. No particular brain regions need to be recorded, and knowledge of the brain regions from which the channels are recording is not required. We also avoid the use of patient specific training data, developing an algorithm that ideally should work adequately in any patient. Additionally, we base the automatic scoring on iEEG data only, with no scalp EEG channels nor EOG/EMG recordings, which represents an immense advantage against how sleep is currently assessed in the epilepsy monitoring unit. The only requirement of the algorithm is that the scoring is carried out on a whole night. In the development of the algorithm, while prioritizing good performance, we also considered interpretability of the algorithm, i.e. the classification should not be a black box. In this way, SleepSEEG could be a tool for studying sleep characteristics in iEEG recordings. Finally, to allow for independent validation and use by the broader community, we made the code for SleepSEEG publicly available at zenodo https://doi.org/10.5281/zenodo.6412063.

2. Methods

2.1. Subjects

We included all patients with drug-resistant focal epilepsy undergoing phase 2 presurgical evaluation at the Montreal Neurological Institute and Hospital between September 2013 and September 2020, who were at least 15 years old, and had a complete night of iEEG recording with stereo-EEG electrodes and simultaneous EOG, EMG, and subdermal scalp EEG electrodes at positions Fz, Cz, Pz, F3, C3, P3, F4, C4, P4, as well as imagining at the time of implantation to determine the location of the electrode positions. This project was approved by the Montreal Neurological Institute and Hospital research ethics board, and all patients gave written informed consent.

From the 55 patients fulfilling the inclusion criteria, 11 were excluded because of poor quality EOG/EMG recordings preventing to score sleep as suggested by the American Academy of Sleep Medicine criteria (Berry *et al* 2017). Note that the

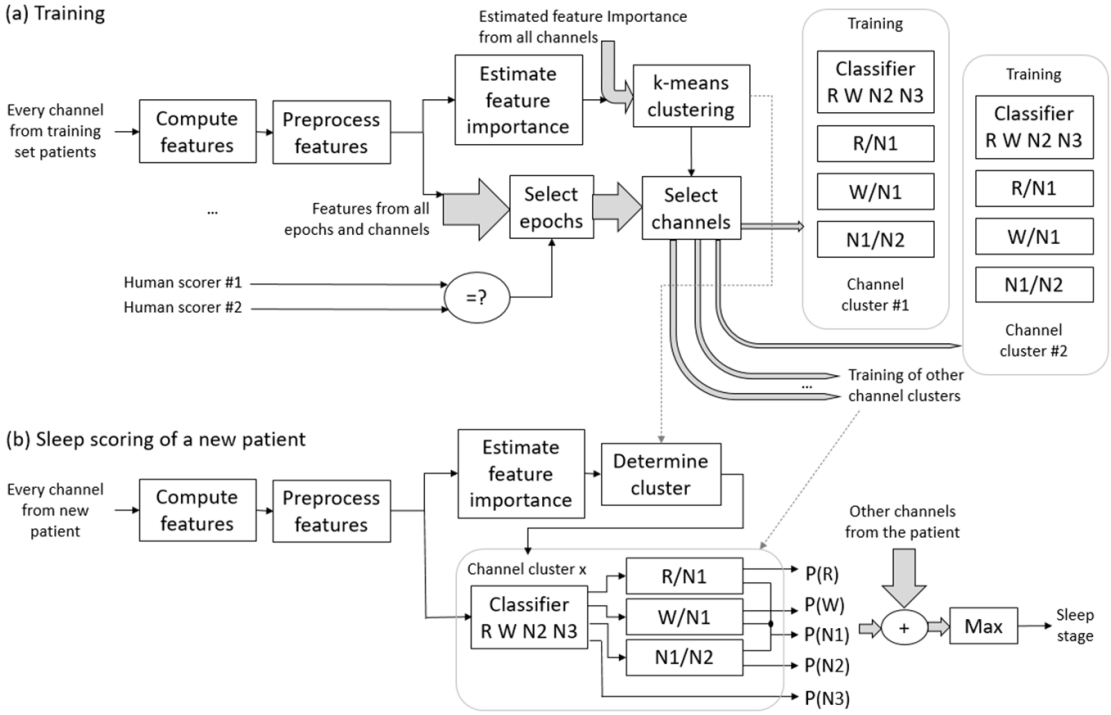


Figure 1. Schematic diagram of the (a) training phase of the algorithm and (b) the scoring of a test patient or a new patient.

automatic scoring would have been possible in these patients, but as the ground truth was not available to assess its performance, these patients were excluded. Sleep scoring of the recordings was carried out independently by two neurophysiologists board certified in sleep and epilepsy trained in different centers. The scoring was carried out between lights-off and lights-on markers. However, differently from the controlled setting in a sleep lab, some patients remained active and awake for up to 4 h after lights-off, while others fell asleep almost immediately. Supplementary table 1 (available online at stacks.iop.org/JNE/19/026057/mmedia) provides clinical information on the patients.

For automatic sleep scoring we used a bipolar montage of neighboring contacts of the stereo-EEG (SEEG) electrodes. The data was acquired at 2000 samples per second, and downsampled to 256 samples per second, the required sampling rate of the automatic scoring code. Note however that the algorithm requires only frequencies up to 64 Hz, meaning that data originally acquired at 200 samples per second can be upsampled to 256 samples per second and used for automatic sleep scoring as well. Other than changing the sampling frequency, no preprocessing was carried out on the EEG signals. The classification algorithm described below handles artifacts and disconnections.

2.2. Development and training

A fourth of the patients were randomly selected as the hold-out set and excluded from the development of the automatic scoring, i.e. feature selection, classification algorithm, and parameter selection. For the remaining patients, a leave-one-patient-out approach was adopted for development of the automatic scoring, with the final training involving all the patients not in the hold-out set. All computations were carried out in Matlab (Mathworks, USA). See figure 1(a) for a schematic of the training process, and figure 1(b) for the automatic scoring process.

2.3. Development strategy

In the development of the automatic scoring algorithm many different preprocessing options, feature combinations, classification approaches, and parameter values were tested. The selection of a particular combination of them was done mainly by checking the median per patient value of the overall agreement with the human scores (percentage of epochs scored as the same sleep stage by any of the human scorers). When the difference between two alternatives was less than 1%, the selection was made by maximizing the minimum per patient agreement, or by selecting the simpler option if there was a difference in complexity between the alternatives.

2.4. Features

We explored 65 features, all based on 30 s intervals corresponding to the sleep scoring epochs of the iEEG signal of a single channel. Explored features included information related to temporal transients (transient increase in power in certain frequency bands with respect to the background in segments of different

length, including the sigma band for spindle detection, and low frequency bands for the detection of slow waves and *K*-complexes, as well as the IED rate), and spectral features exploring both the oscillatory aspects of the spectrum (relative and absolute power in different frequency bands, and its variability in an epoch), and the non-oscillatory characteristics of the spectrum (total power, intercept and slope of the 1/*f* model fitted in different frequency bands, and cumulants of the multifractal spectrum). Finally, we chose a subset of 24 features. The feature selection was based on finding a subset of features that lead to a performance in the development set that was not clearly improved by adding further features (improvement of 1% or less). The feature selection search was not exhaustive, i.e. the search was done in subsets differing in groups of related features (e.g. features corresponding to different frequency bands were either all included or excluded). The final included features are based on a time-frequency decomposition with Daubechies wavelets (8th order) in seven scales or frequency bands (0.5–1–2–4–8–16–32–64 Hz). The magnitude of the coefficients of each channel in each 30 s epoch was used to approximate the power in each band (logarithm of mean value of the squared magnitude of all the coefficients within an epoch in each frequency band), the proportion of power in each band with respect to the total power, and the variability of the power in each band (logarithm of the variance of the coefficients within an epoch). Additionally, we included the first three log-cumulants of the multifractal spectrum that capture non-oscillatory aspects of the spectrum (Wendt *et al* 2007); the log-cumulants are computed based on the wavelet leaders and are related to the 1/*f* exponent, the bandwidth of the validity of the 1/*f* model, and possible existence of a 'knee' in the model. See supplementary table 2 for more information on the selected features.

The preprocessing of the features consisted in excluding outliers by subtracting from the time series a smoothed version of itself (10 sample/5 min moving window average) and excluding the epochs in which median plus/minus 2.5 inter-quartile range (of the whole night) was exceeded. Once the outliers were excluded, we smoothed the feature time-series with a three sample moving window average to minimize the variability. Finally, each time-series was normalized to have zero mean and unit variance. The presence of outliers in the features identifies epochs with artifacts. If the artifacts affect only some channels, the remaining channels can still be used for sleep scoring. If an artifact affects all the channels the classifier assigns the epoch to the W stage, as this is the stage with the overwhelming majority of artifacts. Stage W is also assigned to epochs in which the patient is disconnected from the amplifier, as per the sleep scoring American Academy of Sleep Medicine guidelines (Berry *et al* 2017).

2.5. Unsupervised clustering of channels

iEEG channels record activity from a very localized region around the contacts (von Ellenrieder *et al* 2021), and it was recently shown that different brain regions behave differently during sleep, with some regions undergoing major changes and others only subtle modifications (von Ellenrieder *et al* 2020, Olejarczyk *et al* 2022). Furthermore, channels in any region could be affected by epileptic activity depending on the patient's epileptic network. Therefore, it is unlikely that a single classifier would work well for all channels. Figures 2(a)–(c) shows examples of the time series of a feature in different channels from the same patient, highlighting the difference in local modulation of the feature by the vigilance state. To deal with this issue, we decided to create clusters of channels that show a similar modulation by the vigilance state as a first step before sleep scoring. Note that to compute the channel clusters the sleep stage information should not be used, since the objective is to apply the algorithm to new patients in whom the sleep stage is not yet known. Instead, the clustering of channels was based on the hypothesis that for each channel, the time series of any feature can be thought of as a relatively slow component related to the vigilance state, with added noise in each epoch, as suggested by the examples shown in figure 2. For each channel and feature we computed the proportion of its time series variance explained by a smoothed version of itself (ten sample moving window average), assuming that higher explained variance indicates higher modulation by the vigilance state. This characterizes each channel by 24 values, one for each feature, and we performed an unsupervised clustering with *k*-means algorithm (Euclidean distance, 20 clusters, see supplementary table 2 for range of clustering variables). The resulting clusters contain channels in which each feature is modulated presumably by the vigilance state to the same degree. The proportion of variance explained by the smoothed time series could be interpreted as an estimation of the importance of the feature in sleep scoring, and it can be computed before the actual scoring since it does not depend directly on the sleep stage of the epochs. This relies on our assumption that the major cause of slow fluctuations in the features is relates to vigilance states as it seems to be the case from the examples in figure 2. If this assumption is incorrect, the algorithm would not be able to perform sleep staging.

2.6. Sleep stage classification

Sleep stage classification was performed for each cluster of channels separately. For the channels in each cluster, we used as training data all the epochs that were scored as the same stage by both human scorers. For the classification in one of the five possible sleep stages, we adopted a two-step approach. The first step consisted in a multiclass classification tree with four classes (to discriminate between stages

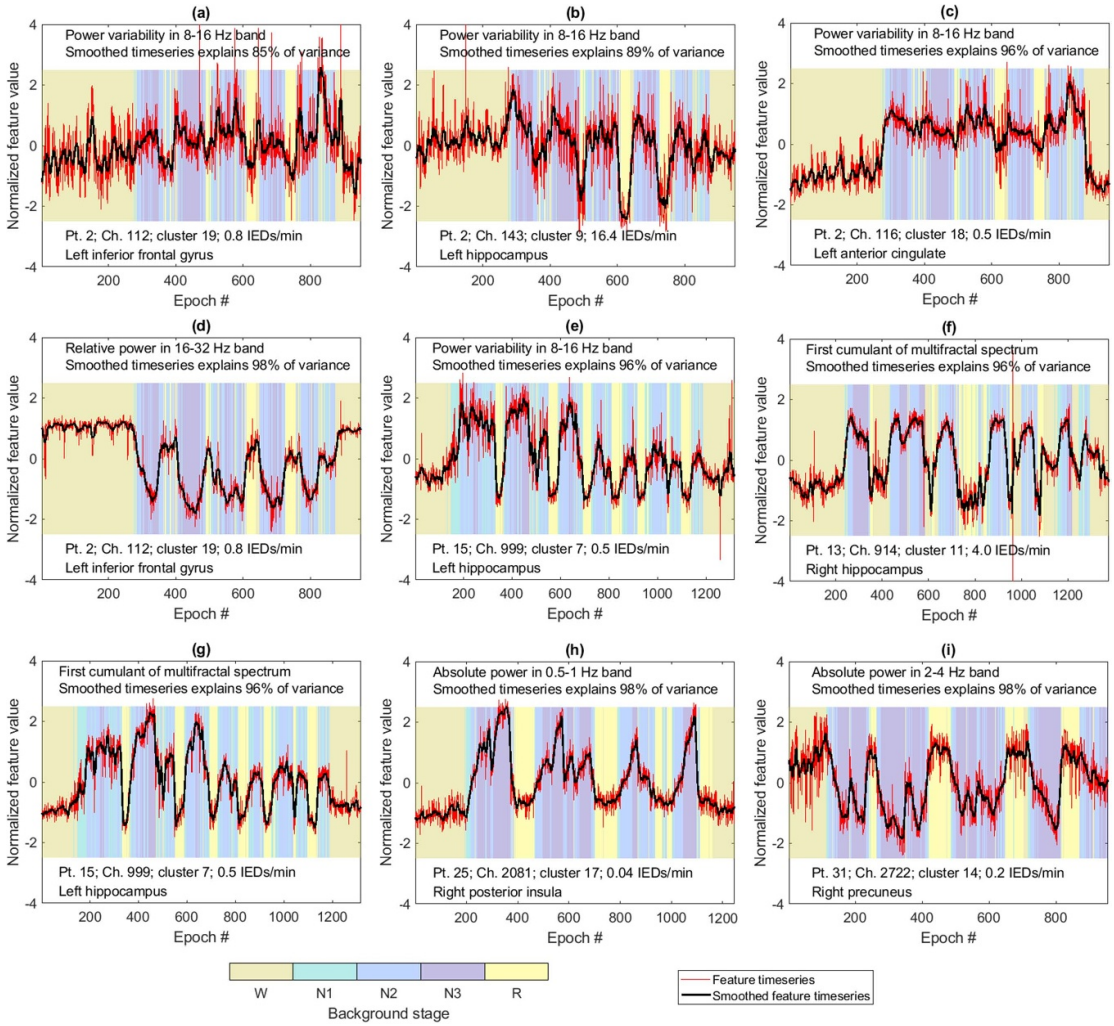


Figure 2. Time series of different features and channels. The red line shows the feature value in each 30 s epoch, the black line is the same time series smoothed in a ten sample (5 min) long moving window. The background color indicates the sleep stage as scored by a human expert. The feature name and percentage of variance explained by smoothed time series are shown on top of the figure, and the patient (Pt.), channel (Ch.), channel cluster number, IEDs rate, and channel location are shown under it. Figures (a)–(c) show different degrees of modulation by sleep of the same feature in different channels of the same subject indicating a variability according to brain region. Figures (d) and (a) show different features on the same channel, in (d) the relative power in the 16–32 Hz band is highly modulated by sleep in this channel suggesting that it is more useful or important for sleep scoring than the power variability in the 8–16 Hz band shown in (a). Figures (e) and (b) show the same feature, power variability in the 8–16 Hz band, in channels recording from the hippocampus in two different patients, in (b) a channel with high IED rate and (e) with low IED rate, note that in addition to different feature values during non-REM sleep, in both cases there is a difference in feature values between stages W and R. Figures (f) and (g) show the same feature, first cumulant of the multifractal spectrum, in channels recording from the hippocampus of two different patients, in (f) a channel with high IED rate and (g) with low IED rate, note that in addition to different feature values during non-REM sleep, in both cases there is a difference in feature values between stages W and R. (h) Absolute power in the 0.5–1 Hz band recording from the insula highly modulated by sleep. (i) Absolute power in the 2–4 Hz band recording from a mesial parietal region highly modulated by sleep.

R, W, N2, N3). To have a balanced training dataset, i.e. a similar number of epochs for the different stages, the maximum number of epochs per stage was limited to 1.5 times the minimum number of epochs per stage in the whole training set. In the final training this meant a maximum number of epochs per stage equal to 1.5 times the number of epochs in stage R. The selection of the excess epochs to exclude for each stage (other than stage R) was done randomly. The second step consisted in three binary classification trees to identify epochs with stage N1. The special treatment of stage N1 was due to its higher scoring difficulty (Rosenberg and van Hout 2013). Following the first step results, the second step attempts to discriminate stage N1 from stage R, W, and N2. Again, the maximum number of epochs per stage was limited to 1.5 times the minimum number of epochs per stage in the training set (in the final training stage N1 was the stage with minimum number of epochs). If the result of the first step was stage N3, we found that the ground truth was very unlikely to be N1, and thus we did not include the second classification step N3/N1 in this case. The final classifier consists then in a multiclass classification tree and three binary trees for each cluster of channels, as depicted in figure 1(a), and is available for download at zenodo https://doi.org/10.5281/zenodo.6412063.

The use of the automatic scoring based on iEEG is necessarily limited to patients with epilepsy undergoing phase 2 presurgical evaluation for epilepsy surgery. For simplicity we included all channels and epochs regardless of the presence of interictal or ictal epileptic activity and refer to this scoring as blind to epileptic activity (BEA). However, if the desired application is to study sleep physiology from direct cortical recordings, epileptic activity could constitute a confounder, i.e. it could influence the values of some features and thus mask the physiological behavior of these features. For this reason, we also trained an automatic scorer that does not rely on channels with abundant interictal activity or epochs during which ictal activity is present. An automatic algorithm was used to detect IEDs (Janca *et al* 2015). Channels with a rate higher than 1 IED min^{-1} were excluded from the training set, as were epochs coinciding with ictal activity (clinical or electrographic seizures as marked by a board-certified neurophysiologist). We identify this automatic scoring as excluding epileptic activity (EEA).

2.7. Testing

Sleep scoring of a recording starts by computing the features for each channel and determining which training channel cluster is the closest to that channel (minimum Euclidean distance to center of clusters). Then, the scoring is done independently for each channel. The result of the channel level scoring is a posterior probability for each stage and epoch, e.g. epoch 410 in channel 52 of patient 35, has posterior probabilities of 17% for stage R, 10% for W, 59% for N2, and 14% for N3. We average with equal weight the posterior probability of all the channels in the recording and select the most likely stage for each epoch (highest averaged posterior probability). If the resulting stage is not N3, a second classification step determines if the stage is N1 or not (with different classifiers based on the result of the first step). Again, this is achieved by averaging with equal weight the posterior probability of stage N1 in all the channels of the patient. Figure 1(b) shows a schematic of this processing.

For determining the performance of the automatic scoring, we used the patients from the hold-out set. To be able to assess the full recording we considered that the automatic scoring of an epoch is correct if it coincides with the scoring of either of the human scorers. Note that restricting the evaluation to epochs scored as the same stage by both human reviewers does not allow to assess all the epochs and excludes the more challenging epochs in which the human scorers disagreed. For each recording we compute the agreement to the human scorers, and the sensitivity and specificity obtained for each stage. High confidence (HC) epochs were defined as epochs in which the automatic scoring algorithm assigned to one of the stages a posterior probability higher than 50%.

2.8. Performance measure and statistical analysis

We report the performance of the automatic scoring as the overall agreement, i.e. number of epochs scored the same by SleepSEEG and either one of the human scores divided by total number of epochs, and by providing the sensitivity and specificity for the different stages (sensitivity computed as number of epochs with correct automatic scoring for each stage divided by the total number of epochs scored as the same stage by either of the human raters, specificity computed as the number of epochs with correct automatic scoring for each stage divided by the total number of epochs scored as the same stage by the algorithm).

We quantify the sleep fragmentation using the sleep stage shift index (SSI), i.e. the number of transitions between different stages per hour, averaged throughout the night. Laffan *et al* (2010), report that in healthy subjects the median (interquartile range) SSI is 10.1 transitions per hour (7.76–13.35). We compute the Pearson correlation between the sleep stage index and the overall agreement (interrater agreement (IR) and agreement between the automatic and human scoring).

To characterize the features during each sleep stage, we compute the mean of the feature value in all channels and all epochs scored as the same stage by both human raters. Note that since the feature values are normalized to zero mean and unit variance for each channel, this mean value is equivalent to the effect size computed as Cohen's *d*. Then, the values can be interpreted in the same scale as Cohen's *d* (0.2 small, 0.5 medium, 0.8 large, 1.2 very large). The effect size of the features in the different sleep stages is an indication of the feature importance for sleep scoring, however, it can only be computed after the scoring is done.

3. Results

3.1. Development and training

The 33 patients of the development set had a total of 3002 channels (1007 frontal, 418 parietal, 102 insula, 1281 temporal, 194 occipital), with a median of 94 (range 40–166) channels per patient, of 9 h 55 min (7 h 55 min–12 h 09 min) recording duration per patient, resulting in a total of 39 550 epochs and 3597 095 epoch-channels. From these, 89% or 35 047 (3199 336) were used for training as both human scorers assigned the same sleep stage. EEA resulted in 1371 channels (605 frontal, 217 parietal, 52 insula, 431 temporal, 66 occipital), 40 (1–115) channels per patient, 39 525 epochs in total and 1646 154 epoch-channels, from which 34 361 (1465 273) were used for training.

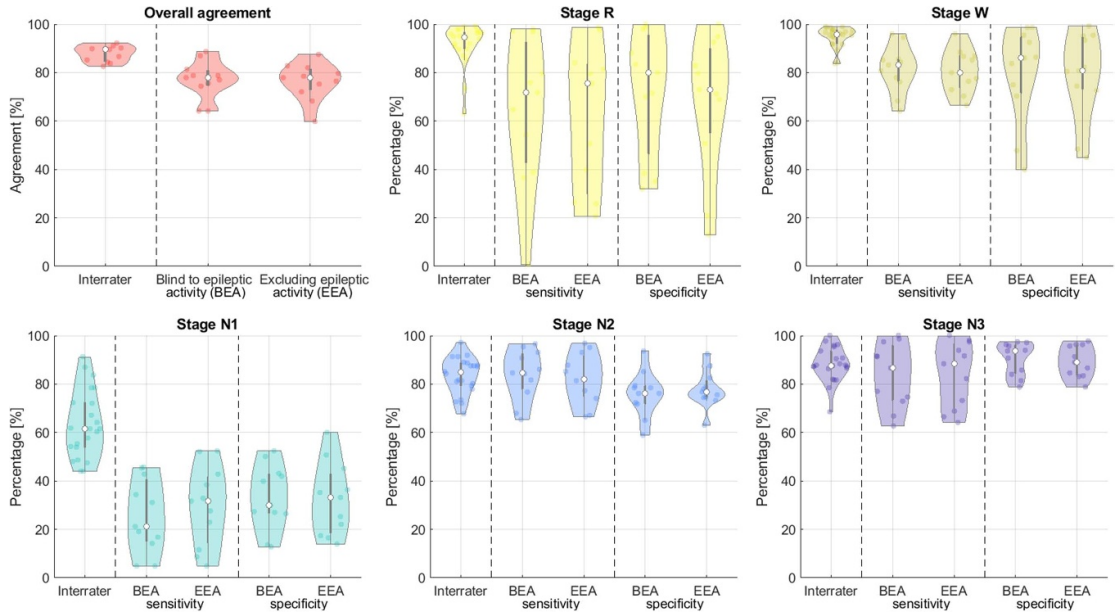


Figure 3. Performance of the automatic scoring. Results are shown for the human IR and both versions of the automatic scoring: BEA and EEA. Note that the human experts scored sleep based on scalp EEG, EOG and EMG, while the automatic scoring is based on iEEG only. Overall agreement is shown in the top left panel, other panels show the sensitivity and specificity of the different stages. The dots within the violins correspond to the performance of the individual patients.

The development and selection of features and parameter values was carried out with a leave-one-patient out approach, i.e. the different methods/parameter values were selected by training with 32 patients and determining the performance in the remaining patient, repeating this 33 times with each patient as testing patient. The approach was used to select the subset of features finally included (absolute and relative power, power variability, and cumulants of multifractal spectrum, all based on wavelet decomposition), the window length of the smoothed time series for the importance estimation and outlier detection (ten epochs), the threshold for outlier exclusion (median $\pm$ 2.5 IQR of the smoothed time series), the window length for feature smoothing (three epochs), the number of channel clusters (20 clusters), the maximum relative number of training epochs per stage (150% of stage with minimum number of epochs), the structure of the classifier (two stages, with the second stage dedicated to identification of N1 stage), the type of classifiers (classification trees, with Gini's diversity index split criterion, 500/50 maximum number of splits for first/second classification step), combination of per channel results into a single per patient result (equal weights of posterior class probabilities). A final training with the 33 subjects was carried out once all parameters had been selected.

3.2. iEEG automatic scoring performance

There were 11 patients in the hold-out set, with a median of 95 (range 40–174) channels per patient. The median duration of the recordings was 9 h 50 min (8 h 29 min–11 h 26 min), with a total of 13 026 epochs and 1249 276 epoch-channels.

The interscorer agreement in the hold-out set was very high. The median overall per-patient agreement was 90% (range 83–92) and Cohen's kappa coefficient was 0.86 (0.77–0.90). Using each of the human raters as ground truth alternately and evaluating the percentage median agreement with the other scorer for each stage we obtained 96% (84–99) for stage W, 62% (44–91) for N1, 85% (68–97) for N2, 88% (69–100) for N3, and 95% (63–99) for R. Figure 3 shows the distribution of the interscorer agreement for all the patients, as well as the results of the automatic scoring.

Note that the interscorer agreement is based on scalp EEG, EOG, and EMG, while the automatic scoring is based only on iEEG and thus a similar performance would be difficult to achieve. The median per-patient overall agreement with the human scorers was 78%, and the median Cohen's kappa coefficient was 0.71. The median per-patient sensitivity of the automatic detector was 83% for stage W, 21% for N1, 85% for N2, 87% for N3, and 72% for R. The observed specificity was 86% for stage W, 30% for N1, 76% for N2, 94% for N3, and 80% for R. Table 1 shows the results including the range among all the patients and figure 3 shows the values for each of the patients. The results show a good balance between specificity and sensitivity, and the best performance is observed for stages N2 and N3. For these stages the performance is similar to the IR. The specificity can be increased by selecting only epochs with posterior probability above 0.5 which we call HC epochs and

Table 1. Automatic scoring performance (based on iEEG only). Median and range among patients in hold-out group are shown.

	BEA automatic scoring			EEA automatic scoring		
	Median	Min	Max	Median	Min	Max
Overall agreement	78	64	89	78	60	88
Cohen's kappa	0.71	0.48	0.84	0.69	0.51	0.85
Stage W sensitivity	83	64	96	80	67	93
Stage N1 sensitivity	21	5	46	32	5	52
Stage N2 sensitivity	85	65	97	82	66	97
Stage N3 sensitivity	87	63	100	89	64	100
Stage R sensitivity	72	1	98	76	21	99
Stage W specificity	86	40	99	81	45	99
Stage N1 specificity	30	13	53	33	14	60
Stage N2 specificity	76	59	94	77	63	93
Stage N3 specificity	94	79	98	89	79	98
Stage R specificity	80	32	100	73	13	100
HC stage W specificity	92	44	100	85	52	100
HC stage N1 specificity	50	0	100	29	0	50
HC stage N2 specificity	82	56	95	81	68	94
HC stage N3 specificity	96	79	98	92	80	98
HC stage R specificity	83	0	100	85	15	100

BEA: blind to epileptic activity, EEA: excluding epileptic activity, HC: high confidence.

constitute 74% (range 36–86) of all epochs. The specificity increased to 92% for stage W, 50% for N1, 82% for N2, 96% for N3, and 83% for R. However, 5/11 patients had less than 5 HC N1 epochs while the rest had none, and 2/11 did not have any stage R HC epochs and 1/11 had less than 5 such epochs.

3.3. Performance excluding epileptic activity (EEA)

Excluding channels with IED rates above 1 min^{-1} and epochs containing ictal activity, the median number of channels per patient was 36 (range 8–90), with a total of 12 519 epochs and 541 205 epoch-channels. The median per-patient overall agreement with either one of the two human scorers was also 78% and the median Cohen's kappa coefficient was 0.69. In this case the median per-patient sensitivity of the automatic detector was 80% for stage W, 32% for N1, 82% for N2, 89% for N3, and 76% for R, see also table 1 and figure 3 for the range and individual patient values. The observed specificity was 81% for stage W, 33% for N1, 77% for N2, 89% for N3, and 73% for R. Again, the sensitivity and specificity are well balanced, and stages N2 and N3 are the most accurately scored. For HC epochs, constituting 77% of all the epochs, specificity increased to 85% or stage W, 29% for N1, 81% for N2, 92% for N3, and 85% for R. Four of eleven patients had less than 5 HC N1 epochs, and 1/11 patients had no HC R epochs while all the rest had more than 5 such epochs. Comparing the results to the previous paragraph, it can be noted that the exclusion of epileptic activity (channels with IEDs and epochs with seizures) does not have an important effect in the scoring performance.

3.4. Confusion matrices

Table 2 shows the confusion matrices for the human interscorer comparison and the automatic scoring approaches. In order to present a summary of all patients, instead of number of epochs which vary among patients, we present the percentage of epochs in each classification class for each approach. Interestingly, the largest number of errors comes from

Table 2. Confusion matrices for human interscorer agreement (based on scalp EEG, EOG, and EMG), and automatic scoring (based on iEEG only). The numbers correspond to percentage of epochs in each patient, in order to combine the different patients. The median and range among all patients is shown.

Human interrater		W	N1	N2	N3	R
	W	33.1 (12.1–43.0)	0.6 (0.1–1.2)	0.3 (0–1.3)	0 (0–0.7)	0 (0–0.4)
	N1	0.6 (0.1–2.0)	4 (2.3–10.3)	0.8 (0.1–2.3)	0 (0.1–0.1)	0.3 (0–1.4)
	N2	0.4 (0.1–1.1)	1.9 (0.5–5.3)	22.8 (16.4–33.3)	1.9 (0.4–3.2)	0.1 (0–1.2)
	N3	0.1 (0–0.2)	0 (0–0.1)	2.9 (0–6.1)	18.6 (11.3–27.9)	0 (0–0)
	R	0.1 (0–0.3)	0.2 (0.1–1.5)	0.2 (0–1.0)	0 (0–0)	11.3 (3.3–19.3)
Automatic scoring BEA		W	N1	N2	N3	R
	W	27.0 (12.0–38.5)	2 (0.2–8.5)	0.5 (0–7.5)	0 (0–0.3)	0.7 (0–10.5)
	N1	1.6 (0–2.1)	1.5 (0.3–3.7)	0.7 (0.3–7.6)	0 (0–0.5)	0.4 (0–6.8)
	N2	1.9 (0.8–4.1)	1.6 (0.3–4.6)	24.7 (16.1–32.7)	2.6 (0–7.7)	0.4 (0–2.9)
	N3	0.1 (0–1.7)	0.1 (0–0.4)	1.4 (0.3–2.0)	15.1 (10.3–22.2)	0 (0–0.2)
	R	1.8 (0–9.2)	0.4 (0–2.9)	0.5 (0–2.3)	0 (0–0)	9.2 (1.0–17.3)
Automatic scoring EEA		W	N1	N2	N3	R
	W	27.5 (11.6–38.5)	2 (0.1–8.5)	0.4 (0–7.5)	0 (0–0.3)	0.1 (0–10.5)
	N1	1.3 (0–2.1)	1.4 (0.4–3.7)	0.7 (0.1–7.6)	0 (0–0.5)	0.3 (0–6.8)
	N2	1.9 (1.0–4.1)	1.9 (0.4–4.6)	24.9 (19.1–32.7)	3.0 (0–7.7)	0.3 (0–2.9)
	N3	0.1 (0–1.7)	0.1 (0–0.4)	1.0 (0.4–2.0)	14.6 (11.9–22.2)	0 (0–0.2)
	R	0.9 (0–9.2)	0.4 (0–2.9)	0.3 (0–2.3)	0 (0–0)	14.6 (0.1–17.3)

the difference in classification of stages N2 and N3 (median of 3%). This is because the number of epochs in these stages is high, and even a good performance leads to a significant number of errors. The largest proportion of errors occurs for stage N1 (against W and N2, median 2%), and in the case of the automatic scoring, for stage R (against W, median 2%) as well.

3.5. Hypnograms

Figure 4 shows the assigned epoch for each of the patients and both human scorers as well as both automatic scoring approaches. Note that in general there is a good concordance, with worse results for patients with more fragmented sleep. There was a significant correlation between the agreement of the automatic scoring and the interscorer agreement (BEA rho $= 0.72$, $p = 0.013$; EEA rho $= 0.79$, $p = .003$). Both the human scorers and the automatic scoring were influenced by the degree of sleep fragmentation; there was a significant anticorrelation between the sleep SSI (averaged from both human scorers) and the automatic scoring performance (BEA rho $= -0.77$, $p = 0.006$; EEA rho $= -0.86$, $p < .001$). In contrast, no statistically significant correlation was observed between the performance of the automatic scoring and the total number of channels or the number of channels in each lobe.

3.6. Features and channel clusters

Figure 5(a) shows the mean values observed for the different features and sleep stages. Note that since the features have been normalized to zero mean and unit variance, the values represent an effect size. These values correspond to all the channel/epochs and give a general idea of which features are of importance in different sleep stages. Figure 5(b) gives the importance (average risk in all splits of the classification tree nodes) of each feature for each channel cluster in the final automatic scoring trained with the 33 patients from the development set. Figure 5(c) compares the feature importance estimated from the modulation of the feature's time series, without knowing the epochs, the importance of the features as the maximum effect size among the different stages, and the maximum feature importance among all channel clusters in the multiclass classifier. Note the similarity between the feature importance estimated without knowing the sleep stages and the importance based on the effect size related to the stages. This indicates that the percentage of variance explained by the smoothed feature time series is an adequate proxy for computing importance. The different features are undoubtedly correlated, e.g. between neighboring frequency bands, the actual classification then only requires a subset of the features as indicated by the sparsity of the importance in the classifier for each cluster (figure 5(b)). In figure 5(a) the mean feature values are very similar between absolute power and power variability. However, there is likely different information in these features since excluding either group leads to a loss in performance in the development patient group.

Figure 5(d) shows the location of the channels in each channel cluster, segregated in different brain regions. As expected, channels in the same brain region tend to be modulated similarly by sleep, and thus a location-based clustering emerges, even though the location was not explicitly used to form the clusters. Clusters have preferential location and preferential features as shown in the figure. The clusters also seem to group channels into either high or low IED rates. This grouping is not perfect, as likely not all channels in a region behave the same, on one hand due to the amount of IEDs, but also due to

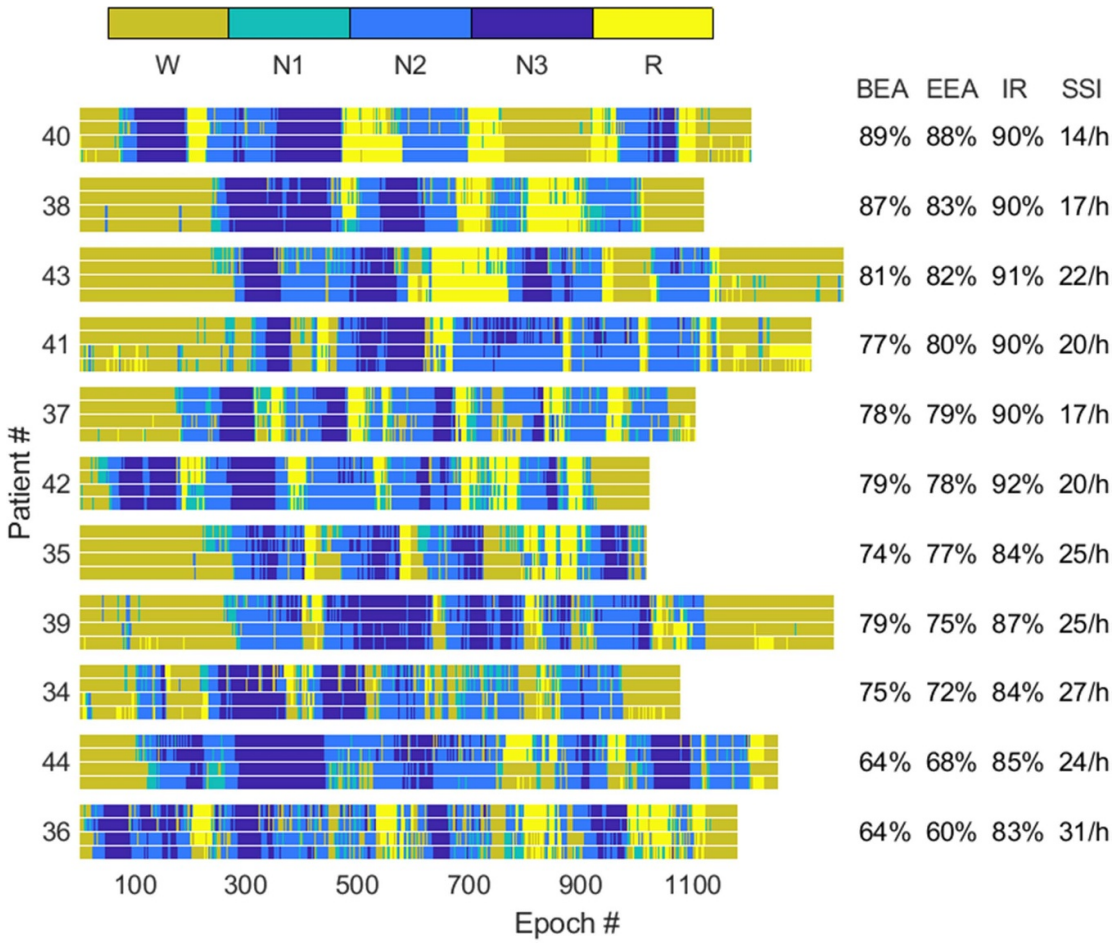


Figure 4. Visual representation of the automatic sleep scoring in the 11 patients of the hold-out set. Each epoch is colored according to the assigned stage. For each patient, the two top rows correspond to the human scorers, the third to the automatic scoring BEA, and the bottom one to the automatic scoring EEA. The patients are sorted in descending overall agreement. The percent agreement for the BEA, EEA, and IR between human scorers are given on the right together with the SSI in number of stage transitions per hour. Note that automatic scoring in patients with more sleep fragmentation tends to perform worse, but the overall sleep structure is preserved except in a few patients (39, 34, 44) in which the automatic scoring of stage R failed.

inter- and intra-patient variability. While the number of clusters was selected through a leave-one-patient-out approach with the training data, it is worth mentioning that 20 clusters produced a result only slightly better than other numbers of the same magnitude, and the resulting clusters are not separated clusters in feature space. However, the clustering does help in the scoring by grouping channels that show the same variability in their features during changes in the vigilance state.

3.7. Examples for study of local sleep

Figure 5 can be used to investigate the local characteristics of sleep and generate hypothesis of interest for further studies. For example, the most useful feature for sleep scoring is the proportion of power in the 16–32 Hz band (mid and high beta band), as shown in figure 5(c). In turn, figure 5(b) shows that this feature is most important in channels from cluster 19, and figure 5(d) shows that this cluster has a large proportion of the channels recording from frontal lateral brain regions. Figure 2(d) shows an example of the time series of the feature in one of these channels, the modulation of the feature by sleep is very clear with almost 99% of the variance explained by the smoothed version of the time series. Note that while it is not surprising that frontal lateral channels can be used to score sleep, the proportion of power in the mid and high beta band is not easily distinguishable by visual analysis in the EEG traces. Figure 5(a) shows that this feature would not be very appropriate for separating stages R and W, nor R and N1.

Figure 5(a) shows that stage R is easiest to distinguish based on the variability of power in the 8–16 Hz band (alpha and low beta), and the first cumulant of the multifractal spectrum. Variability in the 8–16 Hz band is important in channel clusters 18 and 3. Cluster 18 has many channels recording from frontal mesial channels, an example of the feature time series is shown in figure 2(c). Cluster 3 has a large proportion of temporal mesial channels with high IED rate, and figure 2(b) shows the feature time series of such a channel in the hippocampus. While we observe that the feature value is lower during stage R

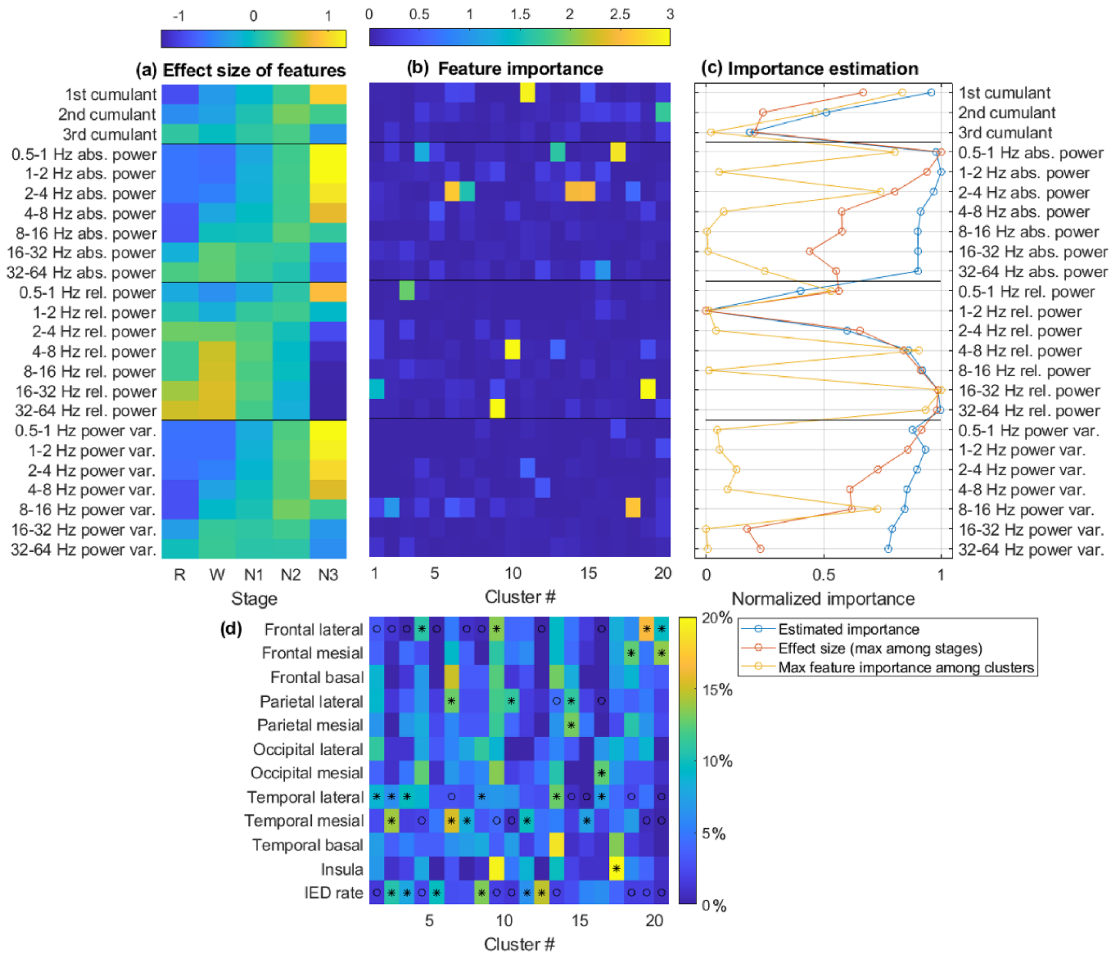


Figure 5. Characterization of the automatic scoring BEA. (a) Mean value of the different features during different sleep stages. The scale can be interpreted as Cohen's *d* effect size (±0.2 small, ±0.5 medium, ±0.8 large, ±1.2 very large). (b) Feature importance for each channel cluster. The sparsity indicates that typically the clusters are organized around a main feature. (c) Estimated importance computed in three different ways: estimated importance based on percentage of explained variance of the feature time series, maximum effect size for the feature among different sleep stages, and maximum importance among the classifiers associated to each channel cluster. The estimate importance was normalized between 0 and 1 for ease of comparison among the three approaches. (d) Percentage of channels from a given brain region in each of the channel clusters, i.e. each row adds up to 100%. A black star (circle) indicates that there are more (less) channels from a region than expected by chance in a given channel cluster. Note that many clusters tend to be organized according to brain region, and most clusters have channels with either low or high IED rate, not a mixture (IED rate indicated in last line).

in channels with high IED rate, it can also be observed in channels with low IED rate as shown in figure 2(e). The other feature important for identifying stage R is the first cumulant of the multifractal spectrum, which is related to the coefficient of the $1/f$ noise model. This feature is of particular importance for cluster 9, which also has many channels recording from temporal mesial regions with high IED rate. An example is shown in figure 2(f), but just as in the case of the 8–16 Hz power variability, the lower feature value during stage R can sometimes be observed in temporal mesial channels with low IED rates as shown in figure 2(g).

Other points to note are the important modulation by sleep stage of the 0.5–1 Hz absolute power in cluster 17, with many channels recording from the insula (example shown in figure 2(h)), and the importance of the 2–4 Hz absolute power in cluster 14, with a high proportion of channels recording from the parietal lobe (example shown in figure 2(i)). For both of these features, figure 5(a) suggest that there would be a poor performance in distinguishing stages R and W. The combination of figures 5(a)–(d) can be used as in the cases mentioned in this section to generate other interesting local sleep hypotheses.

3.8. Automatic scoring excluding epileptic activity (EEA)

The previous paragraphs are based on BEA automatic scoring; however, similar statements would be valid for EEA scoring. Figure 6 shows the feature importance and composition of the channel clusters for EEA scoring. Note that while the clusters are of course different given the lower number of channels, the same sparsity can be observed in figures 5(b), (d), and figures 6(b), (d). Meanwhile, the mean value of the features for each stage is very similar for both scoring approaches, as can be seen by comparing

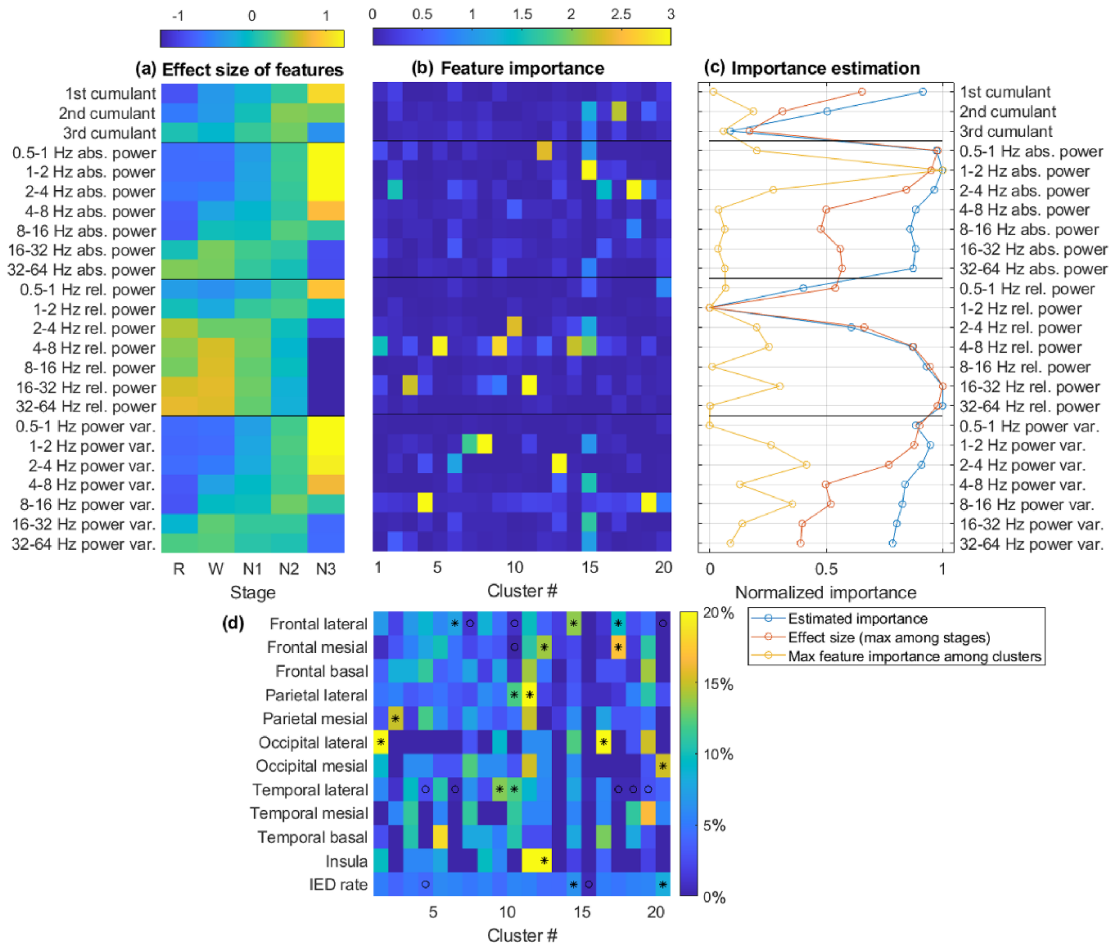


Figure 6. Characterization of the automatic scoring EEA. (a) Mean value of the different features during different sleep stages. The scale can be interpreted as Cohen's *d* effect size (0.2 small, 0.5 medium, 0.8 large, 1.2 very large). (b) Feature importance for each channel cluster. The sparsity indicates that typically the clusters are organized around a main feature. (c) Feature importance computed in three different ways: estimated importance based on percentage of explained variance of the feature time series, maximum effect size for the feature among different sleep stages, and maximum importance among the classifiers associated to each channel cluster. (d) Percentage of channels from a given brain region in each of the channel clusters, i.e. each row adds up to 100%. A black star (circle) indicates that there are more (less) channels from a region than expected by chance in a given channel cluster.

figures 5(a) and 6(a). Figure 7 shows the difference between both approaches (the mean feature values of the BEA minus the EEA approach) that is, it shows the influence of IEDs and seizures in the features. The main changes in mean feature values are an increase in the variability in the 16–32 Hz band power during N3 and a reduction during R, also observed for the 32–64 Hz power variability. The absolute power in the 16–32 Hz band shows an increase during N3, and a decrease in the relative power in the 4–8 Hz band during R. However, all these changes are quite small; the largest is 0.32 and the rest are between 0.23 and 0.2. Most of these changes could be explained by increased IED rates during stage N3 and decreased during stage R.

4. Discussion

Integrating polysomnography in the setting of iEEG is not only challenging but also impractical, and time consuming. To avoid these issues, we developed an automatic sleep scoring algorithm, SleepSEEG, based on iEEG only. The achieved scoring performance is very good, especially considering that the sleep stages are defined based on scalp EEG, chin EMG, and EOG signals, not used by the algorithm. SleepSEEG also returns a confidence value for each epoch, which allows for choosing epochs with a median specificity higher than 80% for all stages except stage N1. We make the code available at zenodo https://doi.org/10.5281/zenodo.6412063 so that it can be easily utilized by the epilepsy and broader neuroscience community. The algorithm is not only useful for scoring sleep during long-term presurgical evaluations, but also provides insights about the local modulation of neuronal activity by sleep in different brain regions. No particular brain regions need to be sampled, and it does not require patient specific training data. The only requirement is that the recording to be scored should encompass a whole night.

Automatic sleep scoring based on non-invasive scalp EEG and/or EOG and EMG is an active research field, both for adults (Fiorillo *et al* 2019, Peter-Derex *et al* 2021) and neonates (Ansari *et al* 2020, Ghimatgar

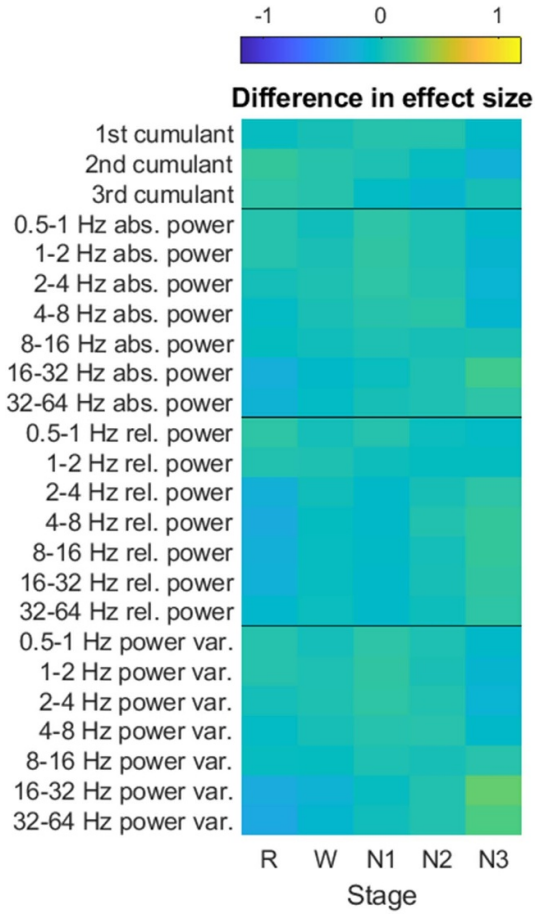


Figure 7. Difference in the mean value of the different features during different sleep stages, shown in figures 5(a) and 6(a), between the scoring BEA and the scoring EEA. The difference would correspond to the effects of high IEDs rate channels (>1 IED min^{-1}) and ictal periods on the features. We kept the same color scale as in figures 5(a) and 6(a), to emphasize that the difference is very small.

et al 2020), and the performance of some of these algorithms can reach the quality of human interrater variability. SleepSEEG is intended for focal epilepsy patients with implanted intracranial electrodes, in whom it is technically challenging or even impossible to acquire standard scalp EEG and EOG/EMG signals, and we do not expect it to achieve the same performance as algorithms that use these standard signals.

4.1. Development of SleepSEEG

Given that the effect of sleep on iEEG is of local character and varies among brain regions (von Ellenrieder *et al* 2020), the automatic scoring relies on multiple classifiers each associated with a cluster of channels. The clusters are obtained in a data-driven approach, and group in the same cluster channels in which the estimated importance of the features is similar. Since this estimated importance needs to be computed for new patients, it cannot use information of the sleep stages. Our approach was to assume that the effect of sleep shows mainly as a slow modulation of the feature, and the performance achieved by the scoring algorithm confirms that it was a valid hypothesis. Thus, features in which a large proportion of the variance of the time series is explained by a smoothed version of the time series were considered important. The resulting clusters were found to include preferentially channels from particular brain regions, and the associated classifiers in many cases depended mostly on one important feature per cluster. We attribute the good performance of the automatic scoring algorithm to this approach as well as to the good quality of the training data, with over 90% median agreement between human scorers trained in different centers, both board certified in sleep and epilepsy and several years of experience as clinical neurophysiologists. The high agreement is especially notable, since sleep scoring is more difficult in neurological patients than in healthy subjects (Norman *et al* 2000, Marzec and Malow 2003, Danker-Hopfe *et al* 2009, Younes *et al* 2016), and several patients in our cohort had very fragmented sleep, a baseline slow wave posterior dominant rhythm, or a diffuse slow wave anomaly. On the other hand, we believe the selection of features was less critical. Features based on wavelet decompositions have been used for sleep scoring with scalp EEG (Sousa *et al* 2015, Peker 2016), and it is a computationally efficient way to obtain different frequency bands. We chose wavelets because they allowed for the computation of both oscillatory and non-oscillatory spectral characteristics, and the selected features showed a good performance in the development set, that was not improved by adding other features. However, they were not necessarily the most informative features when considered individually.

4.2. Epileptic activity did not significantly affect the performance of SleepSEEG

Ignoring or EEA does not affect the performance of SleepSEEG in an important way, suggesting that the features are not very sensitive to the presence of IEDs or to the relatively low proportion of epochs with seizures. Excluding channels with more than 1 IED min^{-1} (more than half of the channels) did not lead to large differences in the estimated importance of the features nor the effect size. This means that SleepSEEG can be used directly in iEEG recordings without worrying about epileptic activity influencing the sleep scoring. However, in cases in which the aim is to study physiological sleep, it is possible to exclude IEDs and seizures, without sacrificing performance of the automatic scoring.

4.3. Identification of stage R was possible without EOG and EMG signals

SleepSEEG performs well for identifying stage R considering that no EOG and EMG signals are available. However, there is a large variability among patients, with very low sensitivity/specificity in a few patients. An interesting finding is that mesiotemporal channels seem to be particularly suitable

for identifying stage R, with lower values of the first cumulant of the multifractal spectrum, and lower power variability in the 8–16 Hz band. The usefulness of the exponent of the $1/f$ model as a feature for distinguishing stage R has been previously reported (e.g. Lendner *et al* 2020), and was the reason for including non-oscillatory features in the detector. The latter could be related to an absence of high amplitude slow waves during REM sleep, which not only has high power but also modulates higher frequency activity; it has been shown that similar to physiological activity (Steriade 2006), epileptic activity is related to sleep slow oscillations (Frauscher *et al* 2015, Nonoda *et al* 2016, von Ellenrieder *et al* 2016, Song *et al* 2017). The variability in power in higher frequency bands is also important, but this is also observed in wakefulness and thus not useful for separating stages W and R.

4.4. N1 was the most challenging stage for both visual and automatic detection

Identification of stage N1 is the most challenging, both automatically and in standard sleep scoring as shown by various authors (Rosenberg and van Hout 2013, Peter-Derex *et al* 2021). In our dataset, the expert scorers achieved a median per patient agreement of 62% for this stage, compared to over 85% for other stages. In automatic scoring based on iEEG the performance was worse, with 21/32% sensitivity and 30/33% specificity for BEA/EEA, even with a second step involving three classification trees specifically trained to detect this stage. The difficulty in staging N1 could reflect a global heterogeneity, difficult to observe in the very local iEEG recordings especially during the wake-to-sleep transition where a high level of asynchrony in falling asleep has been reported between brain regions (Magnin *et al* 2010). In addition, our automatic scoring algorithm does not favor stages that are present for only a few epochs, given the smoothing involved in the estimation of feature importance that is used to cluster the channels. However, it is also possible that the stage is inherently difficult for automatic scoring, since the performance of SleepSEEG is similar to algorithms based on scalp EEG, e.g. ASEEGA achieved 26.5% sensitivity for stage N1 (Peter-Derex *et al* 2021).

The remaining stages, W, N2, and N3, have excellent scoring performance, the median specificity of stage N2 is the only performance measure below 80%, and this is likely due to the low sensitivity of stage N1 and to a lesser degree that of stage R, as some of the epochs in these stages are misclassified as stage N2, thus reducing the specificity. The performance for stage N3 is comparable to the human interrater performance.

4.5. Performance in individual patients

The variability in the automatic sleep scoring performance of different patients correlates with the degree of sleep fragmentation, which also affects the human IR. In other words, sleep is objectively easy or difficult to score in some patients, and we did not identify particular characteristics of the recordings (number or location of channels) that correlated with the automatic scoring performance. The largest inter-patient variability appears in the performance of stage R, and in 1/11 patient the sensitivity to stage R was zero. However, the overall agreement in all the patients of the hold-out group was above 60% (min Cohen's kappa 0.48), i.e. a moderate agreement even for the case with worst performance.

4.6. Comparison of SleepSEEG to existing algorithms for sleep scoring based on iEEG

While there are many examples of automatic sleep scoring algorithms in the literature, most are based on scalp measurements and occasionally EOG/EMG (see e.g. review by Fiorillo *et al* 2019). Since the sleep stages are defined by these measurements it is unreasonable to expect similar performance with intracranial signals only. The situation is further complicated since the patients are at the epilepsy monitoring unit of the hospital, with implanted electrodes connected to the amplifiers, and modifications of brain activity due to the underlying disease. Thus, their sleep is far from normal. Automatic scoring in four epilepsy patients based on subdermal scalp EEG, reached 95% sensitivity and specificity in sleep/wake classification, but this was with an algorithm trained in patient specific data, i.e. the first night was scored and used to train the classifier for the remaining nights (Gangstad *et al* 2019). Reed *et al* (2017) developed a simple classifier to identify slow-wave sleep from intracranial recordings in nine patients, achieving 64% agreement with one human scorer. Automated slow-wave sleep and wakefulness classification in iEEG data was also performed by Kremen *et al* (2017), and they extended their method to differentiate stages W, N2 and N3 in eight implanted epilepsy patients (Kremen *et al* 2019). They showed excellent performance for these stages (87% for N2, 94% for N3), albeit the parameter selection was done with the data of one of the patients later used for testing. The cited study did not attempt to score the more challenging R and N1 stages, although a version with unspecified adaptations including these stages was used by Dell *et al* (2021). Our algorithm includes all sleep stages and achieves an overall good performance.

4.7. Limitations

The main limitation is that the development and testing was done with adult patients from a single center, although further validation with data from several different centers is planned. Since our center uses exclusively depth electrodes in presurgical epilepsy evaluation, this also means that the stated performance is not necessarily guaranteed for grids and strips, nor for pediatric patients.

SleepSEEG requires a full night of recordings to have a complete sample of all the stages and a full range of possible feature values. However, we consider that this is not a limitation from a practical point of view, given the typical duration of presurgical evaluations of a week or more. Note that if a night is available, other segments can be analyzed together with the night, including short naps that do not contain all stages.

4.8. Future work

We plan a validation of the algorithm with external datasets, in order to use it as a standard sleep scoring tool in the Multicenter iEEG Sleep Atlas project (https://IEEGatlas.Loris.ca/) an important ongoing multicenter project. This would also open the possibility of using it in clinical practice for scoring sleep in iEEG presurgical investigations as needed for identifying the most suitable time period for the assessment of interictal biomarkers for identification of the epileptogenic zone (Klimes *et al* 2019, 2022). Furthermore, as shown by a few examples in the manuscript, we plan to use the algorithm to study local sleep through the interpretation of the channel clusters and their relationship to features and brain regions. By doing so, we will be able to have objective and reproducible information on local features of human sleep. Finally, this algorithm could also be easily incorporated in big data analysis necessary for prolonged recordings over weeks, months and years as it is now possible with various neuroresponsive devices, and allow the classification of sleep in relation to circadian, multidien, and perannual rhythms (Baud *et al* 2018, Duun-Henriksen *et al* 2020, Karoly *et al* 2021).

5. Conclusion

We developed SleepSEEG, an automatic sleep scoring algorithm, based on iEEG measurements only, which does not require patient specific training data nor excluding channels with epileptic activity, and achieves a median per patient agreement of 78% with expert human scores. HC epochs in different stages (other than stage N1) can be identified with median per patient specificity above 80%. By considering a trade-off between performance and simplicity at every stage, the automatic scoring algorithm can also be used to gain insight and generate hypotheses regarding local aspects of sleep in different brain regions, and might hence be a valuable tool for sleep evaluation in clinical epileptology and in neuroscience research.

Data availability statement

The data generated and/or analyzed during the current study are not publicly available for legal/ethical reasons but are available from the corresponding author on reasonable request. The code for Sleep SEEG is available for download at 10.5281/zenodo.6412063.

Acknowledgments

This work was funded by project grants of the Canadian Institutes of Health Research (FDN-143208 and PJT-175056) and the Natural Sciences and Engineering Research Council of Canada (RGPIN2020-04127 and RGPAS-2020-00021). B F is supported by a salary award ("Chercheur-boursier clinicien Senior") of the Fonds de Recherche du Québec – Santé 2021-2025.

ORCID iD

Nicolás von Ellenrieder https://orcid.org/0000-0003-0845-347X

References

Ansari A H, de Wel O, Pillay K, Dereymaeker A, Jansen K, van Huffel S, Naulaers G and de Vos M 2020 A convolutional neural network outperforming state-of-the-art sleep staging algorithms for both preterm and term infants *J. Neural Eng.* **17** 016028

Baud M O, Kleen J K, Mirro E A, Andrechak J C, King-Stephens D, Chang E F and Rao V R 2018 Multi-day rhythms modulate seizure risk in epilepsy *Nat. Commun.* **9** 88

Berry R B, Brooks R, Gamaldo C, Harding S M, Lloyd R M, Quan S F, Troester M T and Vaughn B V 2017 AASM scoring manual updates for 2017 (version 2.4) *J. Clin. Sleep Med.* **13** 665–6

Campana C, Zubler F, Gibbs S, de Carli F, Proserpio P, Rubino A, Cossu M, Tassi L, Schindler K and Nobili L 2017 Suppression of interictal spikes during phasic rapid eye movement sleep: a quantitative stereo-electroencephalography study *J. Sleep Res.* **26** 606–13

Danker-Hopfe H *et al* 2009 Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard *J. Sleep Res.* **18** 74–84

Dell K L *et al* 2021 Seizure likelihood varies with day-to-day variations in sleep duration in patients with refractory focal epilepsy: a longitudinal electroencephalography investigation *eClinicalMedicine* **202137** 100934

Dereymaeker A, Pillay K, Vervisch J, de Vos M, van Huffel S, Jansen K and Naulaers G 2017 Review of sleep-EEG in preterm and term neonates *Early Hum. Dev.* **113** 87–103

Duun-Henriksen J, Baud M, Richardson M P, Cook M, Kouvas G, Heasman J M, Friedman D, Peltola J, Zibrandtsen I C and Kjaer T W 2020 A new era in electroencephalographic monitoring? Subscalp devices for ultra-long-term recordings *Epilepsia* **61** 1805–17

Fiorillo L, Puiatti A, Papandrea M, Ratti P-L, Favaro P, Roth C, Bargiotas P, Bassetti C L and Faraci F D 2019 Automated sleep scoring: a review of the latest approaches *Sleep Med. Rev.* **48** 101204

Frauscher B and Gotman J 2019 Sleep, oscillations, interictal discharges, and seizures in human focal epilepsy *Neurobiol. Dis.* **127** 545–53

Frauscher B and Timofeev I 2021 Sleep and seizures *Jasper's Basic Mechanisms of the Epilepsies* 5th edn, ed J L Noebels, M Avoli, M A Rogawski, A Vezzani and A V Delgado-Escueta (Oxford: Oxford University Press) accepted

Frauscher B, von Ellenrieder N, Dolezalova I, Bouhadoun S, Gotman J and Peter-Derex L 2020 Rapid eye movement sleep sawtooth waves are associated with widespread cortical activations *J. Neurosci.* **40** 8900–12

Frauscher B, von Ellenrieder N, Dubeau F and Gotman J 2016 EEG desynchronization during phasic REM sleep suppresses interictal epileptic activity in humans *Epilepsia* **57** 879–88
Frauscher B, von Ellenrieder N, Ferrari-Marinho T, Avoli M, Dubeau F and Gotman J 2015 Facilitation of epileptic activity during sleep is mediated by high amplitude slow waves *Brain* **138** 1629–41
Gangstad S W, Mikkelsen K B, Kidmose P, Tabar Y R, Weisdorf S, Lauritzen M H, Hemmsen M C, Hansen L K, Kjaer T W and Duun-Henriksen J 2019 Automatic sleep stage classification based on subcutaneous EEG in patients with epilepsy *Biomed. Eng. Online* **18** 106
Ghimatgar H, Kazemi K, Helfroush M S, Pillay K, Dereymaker A, Jansen K, Vos M and Aarabi A 2020 Neonatal EEG sleep stage classification based on deep learning and HMM *J. Neural Eng.* **17** 036031
Halasz P, Ujma P P, Fabo D, Bodizs R and Szucs A 2019 Epilepsy as derailment of sleep plastic functions may cause chronic cognitive impairment—a theoretical review *Sleep Med. Rev.* **45** 31–41
Hämäläinen M, Hari R, Ilmoniemi R J, Knuutila J and Lounasmaa O V 1993 Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain *Rev. Mod. Phys.* **65** 413–97
Janca R *et al* 2015 Detection of interictal epileptiform discharges using signal envelope distribution modelling: application to epileptic and non-epileptic intracranial recordings *Brain Topogr.* **28** 172–83
Karoly P J, Rao V R, Gregg N M, Worrell G A, Bernard C, Cook M J and Baud M O 2021 Cycles in epilepsy *Nat. Rev. Neurol.* **17** 267–24
Klimes P, Cimbalnik J, Brazdil M, Hall J, Dubeau F, Gotman J and Frauscher B 2019 NREM sleep is the state of vigilance that best identifies the epileptogenic zone in the interictal electroencephalogram *Epilepsia* **60** 2404–15
Klimes P, Peter-Derex L, Hall J, Dubeau F and Frauscher B 2022 Spatio-temporal spike dynamics predict surgical outcome in adult focal epilepsy *Clin. Neurophysiol.* **134** 88–99
Kremen V, Brinkmann B H, van Gompel J J, Stead M, St Louis E K and Worrell G A 2019 Automated unsupervised behavioral state classification using intracranial electrophysiology *J. Neural Eng.* **16** 026004
Kremen V, Duque J J, Brinkmann B H, Berry B M, Kucewicz M T, Khadjevand F, van Gompel J, Stead M, St Louis E K and Worrell , G A 2017 Behavioral state classification in epileptic brain using intracranial electrophysiology *J. Neural Eng.* **14** 026001
Laffan A, Caffo B, Swihart B and Punjabi N M 2010 Utility of sleep stage transitions in assessing sleep continuity *Sleep* **33** 1681–6
Lambert I, Tramoni-Negre E, Lagarde S, Roehri N, Giusiano B, Trebuchon-Da Fonseca A, Carron R, Benar C-G, Felician O and Bartolomei F 2020 Hippocampal interictal spikes during sleep impact long-term memory consolidation *Ann. Neurol.* **87** 976–87
Lendner J D, Helfrich R F, Mander B A, Romundstad L, Lin J J, Walker M P, Larsson P G and Knight R T 2020 An electrophysiological marker of arousal level in humans *eLife* **9** e55092
Magnin M, Bastuji H, Garcia-Larrea L and Mauguière F 2004 Human thalamic medial pulvinar nucleus is not activated during paradoxical sleep *Cereb. Cortex* **14** 858–62
Magnin M, Rey M, Bastuji H, Guillemant P, Mauguière F and Garcia-Larrea L 2010 Thalamic deactivation at sleep onset precedes that of the cerebral cortex in humans *Proc. Natl Acad. Sci. USA* **107** 3829–33
Marzec M L and Malow B A 2003 Approaches to staging sleep in polysomnographic studies with epileptic activity *Sleep Med.* **4** 409–17
Ng M and Pavlova M 2013 Why are seizures rare in rapid eye movement sleep? Review of the frequency of seizures in different sleep stages *Epilepsy Res. Treat.* **2013** 932790
Nir Y, Staba R J, Andrillon T, Vyazovskiy V V, Cirelli C, Fried I and Tononi G 2011 Regional slow waves and spindles in human sleep *Neuron* **70** 153–69
Nonoda Y, Miyakoshi M, Ojeda A, Makeig S, Juhász C, Sood S and Asano E 2016 Interictal high-frequency oscillations generated by seizure onset and eloquent areas may be differently coupled with different slow waves *Clin. Neurophysiol.* **127** 2489–99
Norman R G, Pal I, Stewart C, Walsleben J A, Rapoport D M 2000 Interobserver agreement among sleep scorers from different centers in a large dataset Sleep **23** 901–8
Olejarczyk E, Gotman J and Frauscher B 2022 Region-specific complexity of the intracranial EEG in the sleeping human brain *Sci. Rep.* **12** 451
Peker M 2016 A new approach for automatic sleep scoring: combining Taguchi based complex-valued neural network and complex wavelet transform *Comput. Methods Programs Biomed.* **129** 203–16
Peter-Derex L, Berthomier C, Taillard J, Berthomier P, Bouet R, Mattout J, Brandewinder M and Bastuji H 2021 Automatic analysis of single-channel sleep EEG in a large spectrum of sleep disorders *J. Clin. Sleep Med.* **17** 393–402
Peter-Derex L, Comte J-C, Mauguière F and Salin P A 2012 Density and frequency caudo-rostral gradients of sleep spindles recorded in the human cortex *Sleep* **35** 69–79
Peter-Derex L, Magnin M and Bastuji H 2015 Heterogeneity of arousals in human sleep: a stereo-electroencephalographic study *NeuroImage* **123** 229–44
Peter-Derex L, Klimes P, Latreille V, Bouhadoun S, Dubeau F and Frauscher B 2020 Sleep disruption in epilepsy: ictal and interictal epileptic activity matter *Ann. Neurol.* **88** 907–20
Reed C M, Birch K G, Kamiński J, Sullivan S, Chung J M, Mamelak A N and Rutishauser U 2017 Automatic detection of periods of slow wave sleep based on intracranial depth electrode recordings *J. Neurosci. Methods* **282** 1–8
Rosenberg R S and van Hout S 2013 The American Academy of Sleep Medicine inter-scorer reliability program: sleep stage scoring *J. Clin. Sleep Med.* **9** 81–87
Sarasso S, Proserpio P, Pigorini A, Moroni F, Ferrara M, de Gennaro L, de Carli F, Lo Russo G, Massimini M and Nobili L 2014 Hippocampal sleep spindles preceding neocortical sleep onset in humans *NeuroImage* **86** 425–32
Song I *et al* 2017 Bimodal coupling of ripples and slower oscillations during sleep in patients with focal epilepsy *Epilepsia* **58** 1972–84
Sousa T, Cruz A, Khalighi S, Pires G and Nunes U 2015 A two-step automatic sleep stage classification method with dubious range detection *Comput. Biol. Med.* **59** 42–53
Steriade M 2006 Grouping of brain rhythms in corticothalamic systems *Neuroscience* **137** 1087–106
Terzaghi M *et al* 2008 Coupling of minor motor events and epileptiform discharges with arousal fluctuations in NFLE *Epilepsia* **49** 670–6
von Ellenrieder N, Frauscher B, Dubeau F and Gotman J 2016 Interaction with slow waves during sleep improves discrimination of physiologic and pathologic high-frequency oscillations (80–500 Hz) *Epilepsia* **57** 869–78
von Ellenrieder N, Gotman J, Zelmann R, Rogers C, Nguyen D K, Kahane P, Dubeau F and Frauscher B 2020 How the human brain sleeps: direct cortical recordings of normal brain activity *Ann. Neurol.* **87** 289–301
von Ellenrieder N, Khoo H M, Dubeau F and Gotman J 2021 What do intracerebral electrodes measure? *Clin. Neurophysiol.* **132** 1105–15
Wendt H, Abry P and Jaffard S 2007 Bootstrap for empirical multifractal analysis *IEEE Signal Process. Mag.* **24** 38–48
Younes M, Raneri J and Hanly P 2016 Staging sleep in polysomnograms: analysis of inter-scorer variability *J. Clin. Sleep Med.* **12** 885–94