Received: 25 October 2017 | Revised: 20 January 2018 | Accepted: 22 January 2018

DOI: 10.1111/jsr.12679

REGULAR RESEARCH PAPER

Journal of Sleep Research

Automatic artefact detection in single-channel sleep EEG recordings

Alexander Malafeev[1,2] | Ximena Omlin[2,3] | Aleksandra Wierzbicka[4] | Adam Wichniak[5] | Wojciech Jernajczyk[4] | Robert Riener[3,6,7] | Peter Achermann[1,2,7]

[1]Institute of Pharmacology and Toxicology, Chronobiology and Sleep Research, University of Zurich, Zurich, Zurich, Switzerland

[2]Neuroscience Center Zurich, University of Zurich and ETH Zurich, Zurich, Switzerland

[3]Sensory-Motor Systems Lab, ETH Zurich, Zurich, Switzerland

[4]Sleep Disorders Center, Department of Clinical Neurophysiology, Institute of Psychiatry and Neurology in Warsaw, Warsaw, Poland

[5]Third Department of Psychiatry and Sleep Disorders Center, Institute of Psychiatry and Neurology in Warsaw, Warsaw, Poland

[6]Medical Faculty, University of Zurich, Zurich, Switzerland

[7]Zurich Center for Interdisciplinary Sleep Research, University of Zurich, Zurich, Zurich, Switzerland

Correspondence
Peter Achermann, University of Zurich, Institute of Pharmacology and Toxicology, Zurich, Switzerland.
Email: acherman@pharma.uzh.ch

Funding information
nano-tera.ch, Grant/Award Number: 20NA21_145929; ETH Zurich Research Grant ETHIIRA, Grant/Award Number: ETH-18 11-1; Swiss National Science Foundation, Grant/Award Number: 32003B_146643, 51NF40-1444639

Summary

Quantitative electroencephalogram analysis (e.g. spectral analysis) has become an important tool in sleep research and sleep medicine. However, reliable results are only obtained if artefacts are removed or excluded. Artefact detection is often performed manually during sleep stage scoring, which is time consuming and prevents application to large datasets. We aimed to test the performance of mostly simple algorithms of artefact detection in polysomnographic recordings, derive optimal parameters and test their generalization capacity. We implemented 14 different artefact detection methods, optimized parameters for derivation C3A2 using receiver operator characteristic curves of 32 recordings, and validated them on 21 recordings of healthy participants and 10 recordings of patients (different laboratory) and considered the methods as generalizable. We also compared average power density spectra with artefacts excluded based on algorithms and expert scoring. Analyses were performed retrospectively. We could reliably identify artefact contaminated epochs in sleep electroencephalogram recordings of two laboratories (healthy participants and patients) reaching good sensitivity (specificity 0.9) with most algorithms. The best performance was obtained using fixed thresholds of the electroencephalogram slope, high-frequency power (25–90 Hz or 45–90 Hz) and residuals of adaptive autoregressive models. Artefacts in electroencephalogram data can be reliably excluded by simple algorithms with good performance, and average electroencephalogram power density spectra with artefact exclusion based on algorithms and manual scoring are very similar in the frequency range relevant for most applications in sleep research and sleep medicine, allowing application to large datasets as needed to address questions related to genetics, epidemiology or precision medicine.

KEYWORDS
computational neuroscience, computerized analysis, electroencephalogram spectral analysis, multiple sleep latency test

1 | INTRODUCTION

Electroencephalographic (EEG) recordings may contain artefacts from many different sources, which is detrimental for quantitative EEG analysis. Thus, artefact detection and exclusion are essential for quantitative EEG analysis. In sleep research, manual marking of artefacts during sleep stage scoring is common, which is time consuming and prevents application to large datasets, i.e. as needed in genetics, epidemiology or precision medicine. Thus, automated methods revealing consistent results are needed. Here we focus on simple

approaches applicable to a single EEG derivation as they should be easily implementable in small portable devices or work on-line and without prior sleep stage scoring.

Technical artefacts, for example powerline noise, may be removed by a band-stop filter (notch filter). However, biological artefacts like muscle, movement, and ocular artefacts and electrical activity of the heart are more difficult to detect as they have a broad variation of appearance.

In general, we have to dissociate between artefact detection (and exclusion for quantitative analyses) and artefact removal ("subtraction" from the EEG). It is difficult to solve artefact subtraction problems exactly. Some signal from the artefact source may remain and part of the useful signal can be removed. It also often requires multiple channels (Delorme & Makeig, 2004; Winkler, Haufe, & Tangermann, 2011), which are not necessarily available with portable devices.

Ocular artefacts can be removed by a number of techniques, for example regression analysis (Semlitsch, Anderer, Schuster, & Presslich, 1986), blind source separation, or independent component analysis (Comon, 1994; Gavelin, Klomp, Priddle, & Uddenfeldt, 2004; Girolami, 1998; Groppe, Makeig, & Kutas, 2009; Lee, Girolami, & Sejnowski, 1999).

Muscle and movement artefacts can tremendously affect the spectra of the EEG recordings, especially in the higher frequency range. These types of artefacts are difficult to detect as they are very variable. However, muscle artefacts have some characteristic properties. Most of the spectral power of a muscle contraction event in the EEG is above 25 Hz (Gotman, Ives, & Gloor, 1981), and muscle artefacts contaminate the high-frequency range (20–80 Hz) with the peak at about 40 Hz, and also affect lower frequencies (Goncharova, Mcfarland, Vaughan, & Wolpaw, 2003). Because muscle artefacts contaminate the higher frequency range, it is possible to apply a low-pass filter (Gevins et al., 1975). However, this may not be the best approach if EEG components above 20–25 Hz are of interest. One of the approaches often applied to avoid problems of filtering or artefact subtraction is the rejection of segments with artefacts. We used this approach and identified 20-s or 30-s EEG segments with artefacts.

We implemented 12 algorithms previously published and developed two new ones (Table 1). Many older papers on artefact detection did not report the performance of the algorithms. We estimated the optimal parameters of the algorithms and evaluated their performance on two types of recordings: nocturnal sleep of healthy participants and patients; and a mixture of sleep and wakefulness in a multiple sleep latency test (MSLT) recorded continuously over approximately 9 hr in patients. Parameter estimation and validation was performed on independent datasets.

TABLE 1 Overview of the applied algorithms and their abbreviations used

Method	Resolution	Online mode possible?
Amplitude thresholding, fixed threshold (ATf)	Sample	Yes
Amplitude thresholding, statistical threshold (ATs)	Sample	No
Slope thresholding, fixed threshold (STf)	Sample	Yes
Slope thresholding, statistical threshold (STs)	Sample	No
Zero crossings (ZC)	Sample	Yes
Mean crossings (MC)	Sample	Yes
Power thresholding 25–90 Hz (PT25)	Sample	Yes
Power thresholding 45–90 Hz (PT45)	Sample	Yes
Power thresholding (average power of epoch) (PTe)	Epoch	Yes
Autoregressive (AR) model	Sample	No
Adaptive AR model, fixed threshold (aARf)	Sample	Yes
Adaptive AR model, statistical threshold (aARs)	Sample	No
K-means (KM) clustering	Epoch	No
Hidden Markov Model (HMM)	Epoch	No

Most of the algorithms return whether a single sample belongs to an artefact or not (resolution sample), but some return whether a whole epoch (20 s or 30 s) contains an artefact or not (resolution epoch). It is also indicated whether an algorithm could be implemented on-line (yes) or whether the entire recording is needed first (no). KM and HMM are the two newly developed algorithms. The algorithms are detailed in Supporting Information.

2 | MATERIALS AND METHODS

2.1 | Datasets

We analysed two datasets, as follows.

(1) Polysomnographic (PSG) recordings of an experiment with vestibular stimulation (Omlin et al., 2018). Three nights (8 hr) of 18 healthy young males (age: 20–28 years; mean: 23.7 years) were recorded: two motion nights (rocking until sleep onset; rocking for first 2 hr after lights out), and a baseline without motion. Recordings included 12 EEG channels, placed according to the 10–20 system, two electrooculogram (EOG) channels, one chin electromyogram (EMG), 1 electrocardiogram (ECG) channel and respiration (chest and abdomen). Data were recorded with the polygraphic amplifier Artisan (Micromed, Mogliano, Veneto, Italy). The signals were sampled at 256 Hz (Rembrandt DataLab; Version 8.0; Embla Systems, Broom Field, CO, USA). Analogue signals were filtered with a high-pass filter (EEG: −3 dB at 0.16 Hz; EMG: 10 Hz; ECG: 1 Hz) and an anti-aliasing low-pass filter (−3 dB at 67.4 Hz). Sleep stages (20-s epochs) were scored according to standardized criteria (Iber, Ancoli-Israel, Chesson, & Quan, 2007). Recordings were performed in the sleep laboratory of the Institute of Pharmacology and Toxicology at the University of Zurich. The Institutional Review Board of the Swiss Federal Institute of Technology in Zurich (ETH Zurich) approved the study. In total, this dataset comprised 53 PSG nighttime recordings of healthy participants.

(2) Polysomnographic data recorded in patients with hypersomnia (two subjects) and narcolepsy (three subjects) who underwent a

MSLT. The EEG was recorded continuously for approximately 9 hr throughout the MSLT. In addition, a night of sleep was recorded in each patient. PSG included six EEG, two EMG, two EOG channels and one ECG. Data were sampled at 200 Hz (polygraphic amplifier, Grass Technologies AURA PSG). Analogue signals were filtered with a high-pass filter (EEG: −3 dB at 0.5 Hz) and an anti-aliasing low-pass filter (−3 dB at 50 Hz). Sleep stages (30-s epochs) were scored according to standardized criteria (Rechtschaffen & Kales, 1968). PSG recordings were performed at the Sleep Disorders Center, Department of Clinical Neurophysiology, Institute of Psychiatry and Neurology in Warsaw, Warsaw, Poland. The Institutional Review Board of Institute of Psychiatry and Neurology approved the study. In total, this dataset comprises five sleep and five MSLT recordings of narcoleptic (n = 3) and hypersomnia patients (n = 2).

To illustrate the sleep structure in the EEG and the occurrence of large artefacts, spectrograms were calculated. Power density spectra were determined for 20-s or 30-s epochs (FFT; average of five 4-s or six 5-s epochs without overlap; Hanning window). Spectra are plotted and colour-coded on a logarithmic scale (spectrograms, Figures 1–3).

Artefacts were visually scored by experts in both datasets on an epoch basis (i.e. each 20-s or 30-s epoch has a label whether it contains an artefact or not). Please note that in dataset 1 (healthy subjects), only severe artefacts were visually identified. Scorers were instructed to mark only severe artefacts. Afterwards a semiautomatic procedure based on EEG power in the 20–40 and 0.75–4.5 Hz range was applied to detect small artefacts. Artefact markings of the semiautomatic procedure were used only in the original study (Omlin et al., 2018). Artefacts in the second dataset (patients) were scored by the first author and all artefacts were marked. This might explain why the FPRs in the second dataset were smaller than in the test data of the first dataset (Table 2).

We analysed derivation C3A2 in the context of this paper. The parameters of the algorithms are to some degree dependent on the

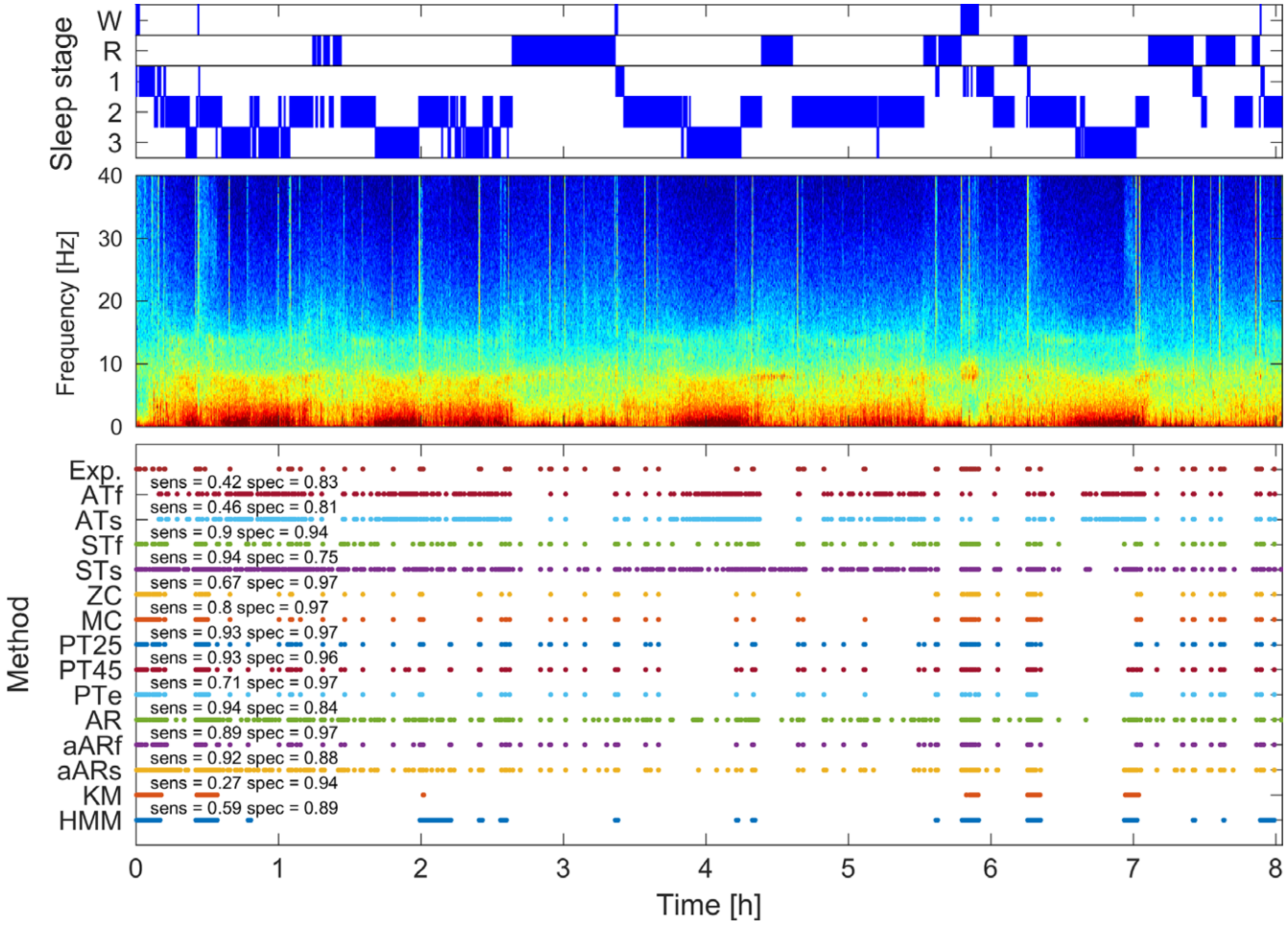

FIGURE 1 Example of artefact detection in a recording of dataset 1. Top: hypnogram (W, waking; R, rapid eye movement [REM] sleep; 1–3, non-rapid eye movement [NREM] sleep stages N1–N3). Middle: spectrogram (power density spectra of 20-s epochs colour-coded on a logarithmic scale [0 dB = 1 μV^2/Hz; −10 dB 20 dB]). Bottom: artefacts marked by an expert (Exp.) and artefacts determined by the different algorithms (see Table 1 for meaning of abbreviations). Dots corresponding to 20-s epochs marked as an artefact. Note that due to the condensed display, dots may overlap. Sensitivity (sens, true-positive rate [TPR]) and specificity (spec, 1 − false-positive rate [FPR]) achieved by the different algorithms are indicated

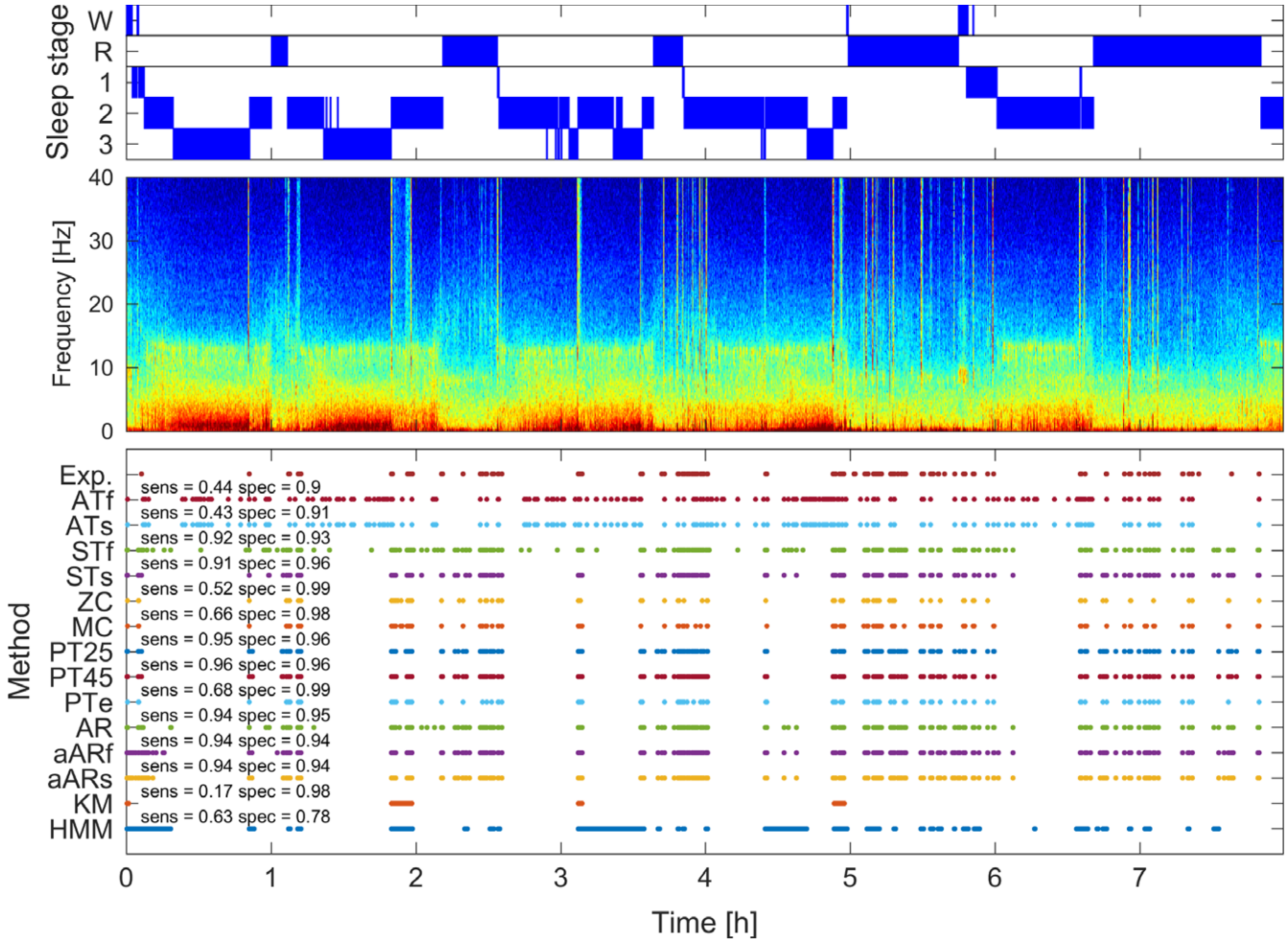

FIGURE 2 Further example of artefact detection in a recording of dataset 1. For details, see Figure 1

derivation used. If referential (mastoid reference) derivations (frontal, central, occipital) as classical for sleep recordings are used we do not expect that adaptations are needed. However, working with, for example, bipolar recordings would require adaptation of the parameters.

We mainly focused on simple algorithms that are easy to implement and were used in the past. Two additional algorithms were developed and tested. In contrast to the other algorithms, these two have no tunable parameters as they cluster the data in two categories (no artefacts, artefacts; see Supporting Information).

Most of these algorithms produce a classification for each sample (Table 1), i.e. an outcome of an algorithm is an array with labels for each sample whether it belongs to an artefact or not. We translated this information into an epoch-wise classification in order to be able to compare the outcome of an algorithm with our expert classification. We classified an epoch as an artefact if it contained at least one sample identified as an artefact.

2.2 | Algorithms

The implemented algorithms (Table 1; abbreviations) and the corresponding parameters (derived and applied are listed in Table 2) are described in the Supporting Information.

We do not expect a noticeable influence of the sampling rates and filter settings of the recording equipment used because the algorithms were chosen to work in the frequency range of 0.5–90 Hz. This is evident as parameters derived on healthy participants were transferable to patients recorded with a different system. However, for sampling rates <200 Hz adaptations would be needed.

2.3 | Evaluation of the performance of the algorithms

To evaluate the performance of the algorithms, we randomly split the data of the first dataset (sleep data) into a training and testing set in proportion of 60%–40% (32 and 21 recordings). We computed receiver operator characteristic (ROC) curves (Green & Swets, 1966) for each algorithm and recording of the training dataset. To compute the ROC curves, parameters of the algorithms (thresholds) were systematically varied in a certain range (Table 2). ROC curves are plots that give an understanding about the performance of a binary classifier. On the *x*-axis, the FPR (percentage of clean epochs marked as containing artefacts) is plotted; and on the *y*-axis, the true-positive rate (TPR; percentage of epochs with artefacts that were marked as

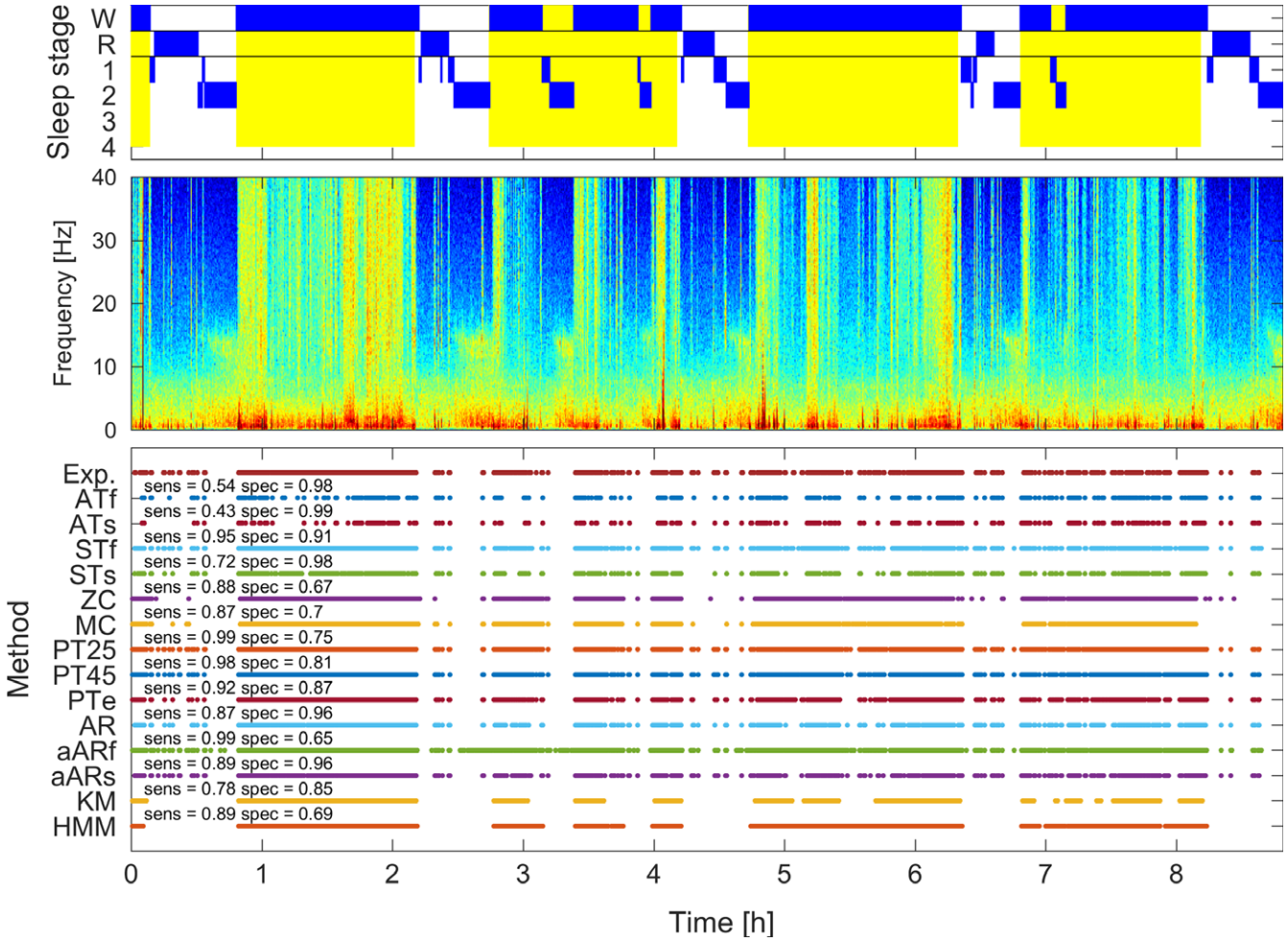

FIGURE 3 Example of artefact detection in a multiple sleep latency test (MSLT) recording of dataset 2. Top: hypnogram (W, waking; R, rapid eye movement [REM] sleep; 1–4, non-rapid eye movement [NREM] sleep stages 1–4). Yellow areas indicate lights on. Middle: spectrogram (power density spectra of 30-s epochs colour-coded on a logarithmic scale [0 dB = 1 $\mu V^2/Hz$; −10 dB 20 dB]). Bottom: artefacts marked by an expert (Exp.) and artefacts determined by the different algorithms (see Table 1 for meaning of abbreviations). Dots corresponding to 30-s epochs marked as an artefact. Note that due to the condensed display, dots may overlap. Sensitivity (sens, true-positive rate [TPR]) and specificity (spec, 1 − false-positive rate [FPR]) achieved by the different algorithms are indicated

having artefacts) is plotted. One varies the threshold and calculates FPR and TPR for each value of a threshold. Plotting of these points forms the ROC curve. In case of random results, the ROC curve would be a straight line with the area under the curve (AUC) equal to 0.5. The AUC is a marker of the quality of an algorithm; the larger the AUC, the better the performance of the algorithm. ROC curves for ATf, STf and PT25 are illustrated in Figure 4.

There is no single way to choose an optimal threshold for the application of the algorithms, as sensitivity and specificity cannot be increased concomitantly (Habibzadeh, Habibzadeh, & Yadollahie, 2016). Moreover, it is application dependent whether priority is given to sensitivity or specificity. In some applications the "costs" of false-negatives are large; in other applications the "costs" of false-positives are large. In our case, false-negatives may distort average spectra in the frequency range of interest, whereas exclusion of some additional clean epochs (false-positives) would not affect average spectra. After visual inspection of the ROC curves, we decided to apply a threshold corresponding to a false-positive rate close to 0.1 (not based on optimization). Our choice of a FPR fixed at 0.1 has consequences, i.e. it leads to the rejection of 10% of all epochs not marked as artefactual by the expert to be rejected. As, however, experts only marked severe artefacts (dataset 1, see above), we consider this as justifiable. Experts excluded 7 ± 3% of epochs. Therefore, the total rejected epochs should theoretically fluctuate at about 16.3%.

For the test datasets, we applied those thresholds and computed specificity (percentage of artefact-free epochs correctly marked as artefact free) and sensitivity (equal to TPR) and compared resulting performance measures with performance on the training set, and average power density spectra of non-rapid eye movement (NREM) sleep were calculated to assess the impact of artefact exclusion. Additionally, we applied the algorithms with the same thresholds to patient data of the second dataset and validated performance additionally on sleep and MSLT recordings.

TABLE 2 Parameters and performance of the algorithms

	Dataset 1 (training data)			Dataset 1 (test data)		Dataset 2 (MSLT data)		Dataset 2 (sleep data)	
	AUC (std)	Thres	TPR (std)	TPR (std)	FPR (std)	TPR (std)	FPR (std)	TPR (std)	FPR (std)
Amplitude thresholding, fixed threshold (ATf)	0.820 (0.059)	168.333 μV	0.556 (0.097)	0.496 (0.122)	0.097 (0.062)	0.514 (0.262)	0.061 (0.103)	0.405 (0.128)	0.014 (0.024)
Amplitude thresholding, statistical threshold (ATs)	0.819 (0.059)	5.292 σ	0.489 (0.181)	0.420 (0.151)	0.087 (0.075)	0.310 (0.159)	0.026 (0.033)	0.398 (0.153)	0.014 (0.023)
Slope thresholding, fixed threshold (STf)	0.953 (0.023)	928,336.64 μV/s	0.905 (0.051)	0.886 (0.067)	0.103 (0.067)	0.946 (0.035)	0.143 (0.100)	0.696 (0.140)	0.043 (0.052)
Slope thresholding, statistical threshold (STs)	0.952 (0.023)	3.700 σ	0.905 (0.041)	0.868 (0.137)	0.138 (0.093)	0.763 (0.136)	0.076 (0.086)	0.726 (0.232)	0.055 (0.052)
Zero crossings (ZC)	0.866 (0.068)	36.250 # per s	0.692 (0.126)	0.644 (0.105)	0.072 (0.041)	0.949 (0.045)	0.338 (0.071)	0.727 (0.109)	0.169 (0.263)
Mean crossings (MC)	0.917 (0.047)	40.000 # per s	0.791 (0.100)	0.756 (0.074)	0.079 (0.042)	0.951 (0.047)	0.273 (0.084)	0.720 (0.100)	0.131 (0.236)
Power thresholding 25–90 Hz (PT25)	0.966 (0.028)	8.400 μV^2	0.923 (0.058)	0.905 (0.079)	0.108 (0.078)	0.994 (0.002)	0.245 (0.107)	0.902 (0.072)	0.076 (0.075)
Power thresholding 45–90 Hz (PT45)	0.962 (0.031)	3.143 μV^2	0.909 (0.073)	0.899 (0.083)	0.097 (0.069)	0.988 (0.005)	0.206 (0.093)	0.855 (0.094)	0.045 (0.052)
Power thresholding (average power of epoch) (PTe)	0.926 (0.037)	5.263 μV^2	0.780 (0.073)	0.749 (0.075)	0.091 (0.083)	0.970 (0.028)	0.228 (0.125)	0.833 (0.017)	0.059 (0.069)
Autoregressive model (AR)	0.954 (0.023)	3.458 σ	0.911 (0.046)	0.878 (0.121)	0.137 (0.084)	0.865 (0.104)	0.109 (0.111)	0.820 (0.252)	0.176 (0.166)
Adaptive AR, fixed threshold (aARf)	0.956 (0.021)	11.111 μV	0.897 (0.065)	0.887 (0.071)	0.111 (0.080)	0.990 (0.004)	0.326 (0.123)	0.922 (0.052)	0.134 (0.102)
Adaptive AR, statistical threshold (aARs)	0.956 (0.021)	3.333 σ	0.906 (0.050)	0.884 (0.113)	0.138 (0.074)	0.881 (0.084)	0.111 (0.111)	0.823 (0.256)	0.121 (0.090)
K-means (KM)				0.406 (0.151)	0.190 (0.118)	0.902 (0.077)	0.183 (0.088)	0.346 (0.170)	0.050 (0.041)
Hidden Markov Model (HMM)				0.652 (0.176)	0.269 (0.135)	0.924 (0.025)	0.216 (0.083)	0.754 (0.179)	0.368 (0.185)

The first three columns correspond to the training of the algorithms. Area under the curve (AUC) with its standard deviation, optimal threshold (Thres; σ, standard deviation), true-positive rate (TPR) at a false-positive rate (FPR) close to 0.1. Columns 4 and 5 represent the TPR and FPR obtained by applying the algorithms to the test data of dataset 1. The further columns represent TPR and FPR obtained by applying the algorithms to the multiple sleep latency test (MSLT) and sleep data of dataset 2 (patients). K-means (KM) clustering and hidden Markov models (HMM) did not work on sleep data of datasets 1 and 2, which contained a very small number of epochs with artefacts, which is insufficient for these unsupervised classifiers to learn that artefacts are a separate class.

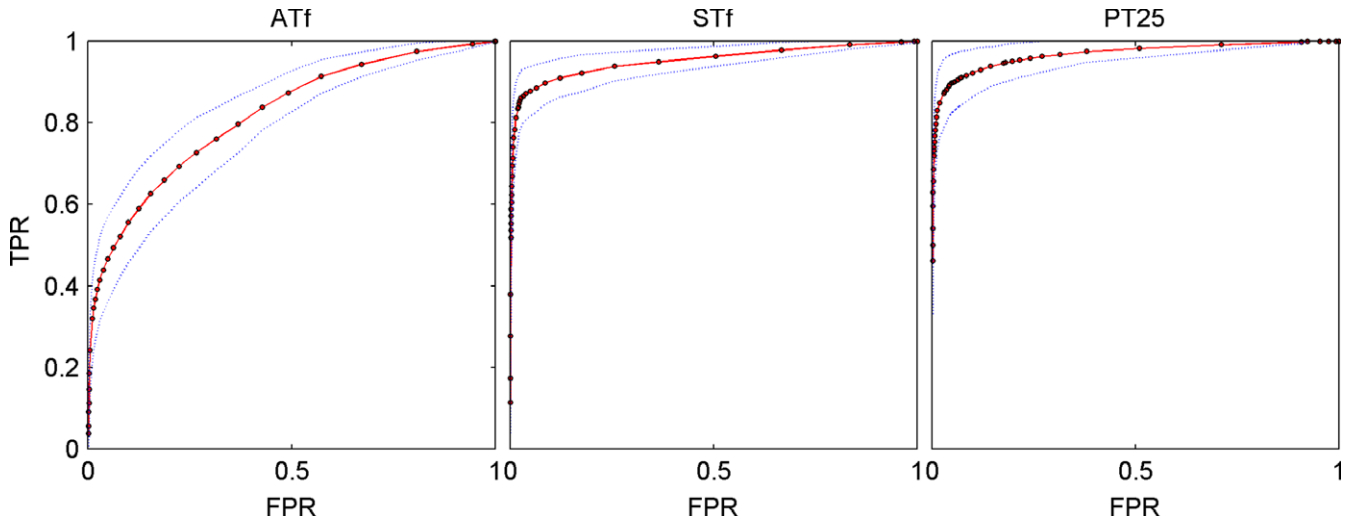

FIGURE 4 Average receiver operating characteristic (ROC) curves (mean and standard deviation across training set) for the algorithms "Amplitude thresholding, fixed threshold ATf" (left), "Slope thresholding, fixed threshold STf" (middle) and "Power thresholding 25–90 Hz PT25" (right). Dots with the red line show average ROC curve among recordings in the training set. FPR, false-positive rate; TPR, true-positive rate. Blue curves depict standard deviations

3 | RESULTS

3.1 | Derivation of parameters (thresholds) of the algorithms

Areas under the ROC curves, optimal thresholds (we chose them in a way that FPR ~ 0.1) and TPR resulting from the training datasets are depicted in Table 2 (columns 2–4). Seven algorithms showed quite good performance (AUC > 0.95), with PT25 showing the best performance, i.e. largest AUC and TPR.

3.2 | Testing of performance on independent datasets

The performance of the algorithms was tested on the test data of dataset 1 and data (MSLT and sleep) of dataset 2 applying the derived thresholds.

Performance of the algorithms (Table 2, columns 5 and 6), i.e. the TPR, was somewhat lower for the test data than for the training data. PT25 was again performing best. Performance of algorithms KM and HMM was not satisfactory (Table 2). Examples of artefact detection applied to single-sleep recordings of dataset 1 are illustrated in Figures 1 and 2. Note that the performance of some algorithms varies considerably from recording to recording (e.g. STs or AR in Figures 1 and 2). The values of TPR and FPR reported in Table 2 are average values. The fluctuations in the performance of the algorithms applied to different recordings are reflected in the standard deviations shown in brackets.

Performance of the algorithms applied to patient data of a different laboratory (dataset 2) was also good (TPR ≥ 0.9; FPR ~ 0.1; Table 2, columns 7–10; sensitivity = TPR; specificity = 1 − FPR), thus the determined thresholds are generalizable. KM and HMM performed well on the MSLT data with a lot of intermittent wakefulness (Table 2, columns 7 and 8). However, performance on sleep data (Table 2, columns 9 and 10) was not satisfactory. Figure 3 illustrates artefact detection in a MSLT recording of dataset 2.

3.3 | Effect of artefact exclusion on NREM sleep power density spectra

An important purpose of artefact identification is to be able to obtain clean average power density spectra for further evaluation. Figure 5 (top rows) illustrates how artefact removal with ATf, STf and PT25, and by an expert affected average NREM sleep power density spectra of a single subject. Artefact removal affected mainly frequencies above 16 Hz. The algorithms excluded generally more epochs (approximately twice as many) than an expert, as the parameters were derived with a FPR set at 10%. In the case of artefact exclusion with ATf, STf and PT25, variability of the average spectra (standard deviation) was smaller than after artefact exclusion by experts (Figure 5). Artefact exclusion mainly resulted in reduced power density in frequencies above 16 Hz. With ATf, less high-frequency artefacts were removed than by expert scoring (Figure 5, bottom, red curve above green one), while with STf and PT25 more high-frequency artefacts were removed than by the expert marking. However, there was no difference in frequencies below 16 Hz. Due to the large inter-individual differences, we were not able to find statistically significant differences in average power density spectra between no artefact exclusion and artefact exclusion by algorithms or an expert. How close average spectra match between expert marking and algorithms may be another benchmark to assess the quality of an algorithm.

4 | DISCUSSION

We performed a systematic evaluation of mostly simple algorithms that can easily be implemented and demonstrated that they work

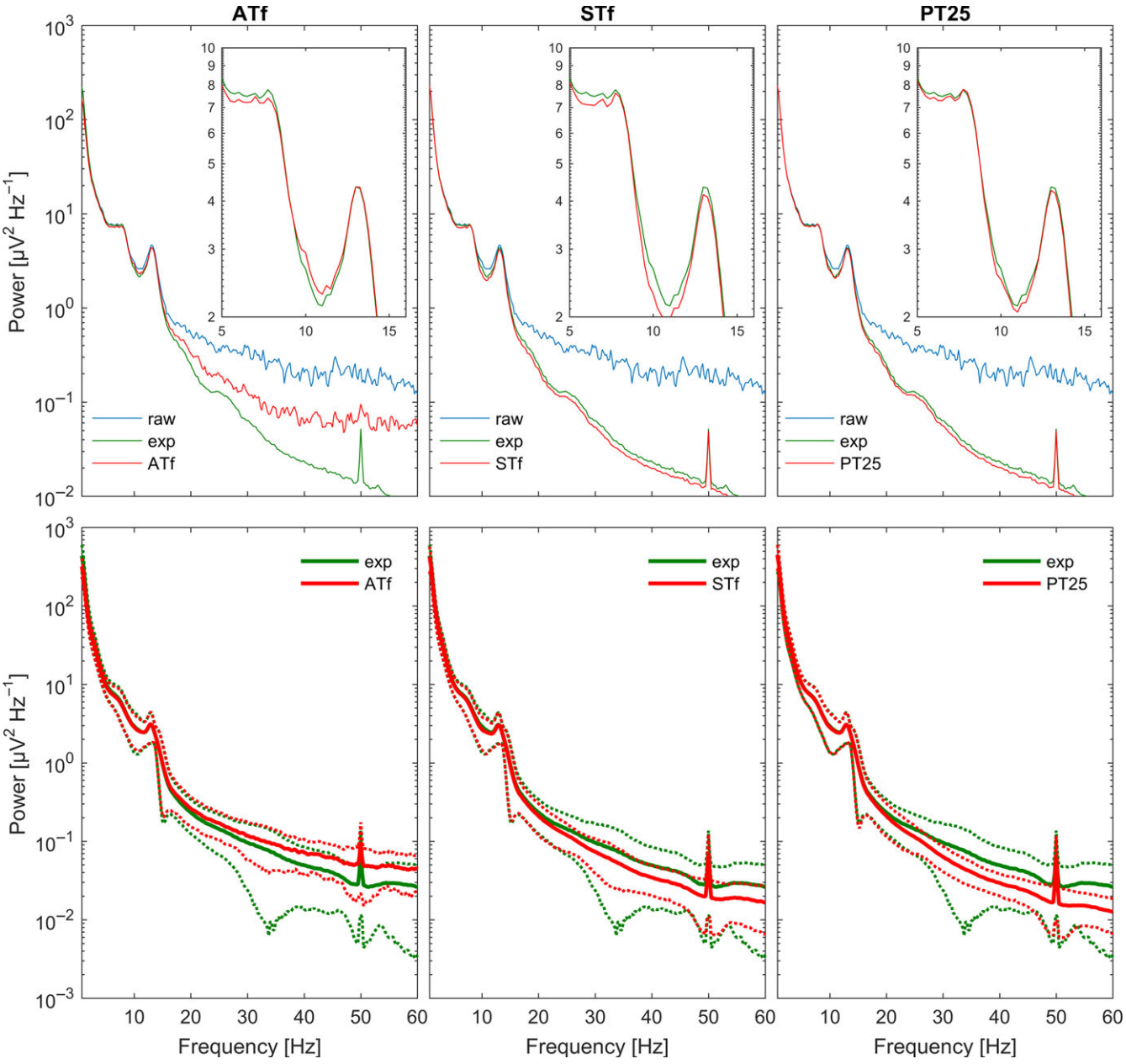

FIGURE 5 Impact of artefact exclusion on average power density spectra. Top rows: average power density spectra of non-rapid eye movement (NREM) sleep of 1 night of a single subject. No artefact exclusion (blue), artefacts excluded by an expert (green), and artefacts excluded by algorithms (red). Inset illustrates spectra between 5 and 16 Hz. Bottom rows: average power density spectra of NREM sleep across subjects (n = 21) of the test dataset. Dashed lines show standard deviations. Only spectra after artefact removal are shown. First column: amplitude thresholding, fixed threshold (ATf). Second column: slope thresholding, fixed threshold (STf). Third column: power thresholding 25–90 Hz (PT25). With ATf less high-frequency artefacts were removed than by expert scoring (red curve above green one), while the two other methods removed more high-frequency artefacts than by expert scoring

well reaching moderate to good sensitivity (TPR) while specificity (1 − FPR) was fixed at 0.9. A recent paper of a specific algorithm reported specificity of approximately 0.95 (D'Rozario et al., 2015), and Durka, Klekowicz, Blinowska, Szelenberger, and Niemcewicz (2003) observed FPRs ranging from 0.04 to 0.14 between different raters, and of 0.06 and 0.08 in one rater who scored artefacts twice at an interval of 3 weeks. The evaluated methods showed good precision to obtain clean average power density spectra as an example of quantitative EEG analyses. Our aim was not to identify particular types of artefacts like, for example, contamination by eye movements, but to establish a reliable procedure to exclude artefacts to be able to obtain reproducible clean quantitative EEG measures as, for example, mean power density spectra, circumventing manual artefact scoring, which is time consuming and to some degree subjective (Anderer et al., 1999; Coppieters't Wallant et al., 2016). Many previous papers focused on a specific algorithm (Coppieters't Wallant et al., 2016; D'Rozario et al., 2015) or reviewed approaches more generally, not assessing their performance or did not provide

parameters that could be applied (Barlow, 1983, 1984, 1986; Bodenstein & Praetorius, 1977; Durka et al., 2003; Gotman et al., 1981; Ktonas, Osorio, & Everett, 1979).

The estimated thresholds (parameters) of the algorithms (provided in Table 2) were robust and did not suffer from overfitting as the results were not specific for the dataset used for parameter estimation. Overfitting is a phenomenon when an algorithm learns properties of a specific subset of the data and, despite the excellent performance on the training data, it shows bad performance on the new data. The tradeoff between the number of parameters and quality of the fit is called bias-variance tradeoff (Geman, Bienenstock, & Doursat, 1992), and should be taken into account. The thresholds could be applied to independent datasets and data of another laboratory with different types of recordings (sleep and MSLT) reaching the same performance as with the training dataset. However, we studied adult EEG data, and in particular derivation C3A2. Thus, for different derivations or applications in children or infants the thresholds may have to be adapted, in particular for amplitude and slope thresholding.

As we demonstrated, even very simple methods can provide a good performance that is suited for practical applications. However, in the context of a particular application, a tradeoff between excluding too much data and not excluding enough artefacts needs to be found. For example, algorithms that capture high-frequency features showed the best performance. However, for sleep applications focusing on the slow wave or spindle frequency range, these algorithms might exclude too much of the data as many artefacts mainly affect power density spectra above 20 Hz (Goncharova et al., 2003; Figure 5). For such applications, we recommend decreasing the sensitivity of the algorithms. Additionally, using a combination of different features to detect artefacts may improve the performance (Coppieters't Wallant et al., 2016).

We developed additionally two non-supervised methods for artefact detection, which did not require predefined parameters. For this purpose, we employed HMM and K-means clustering to dissociate clean EEG from artefacts. However, these two algorithms worked well only with enough wake (artefacts due to movements) and sleep data as in the case of continuous MSLT recordings over 9 hr. If only a small percentage of the recording is contaminated by artefacts as in the night recordings, the clustering turned out to be not reliable.

We excluded entire scoring epochs (20 s or 30 s) whenever an artefact was detected. For sleep EEG recordings, this approach leaves enough data for subsequent analyses. However, this might not be the case for shorter wake EEG recordings. In general, the algorithms work equally well with shorter epochs and can thus be easily adapted to the needs of the analyses. However, an important finding was that it is preferable to detect artefacts with a high temporal resolution: if we compute, for example, power in the high-frequency range averaged across an epoch, the power of an epoch with an artefact will not be very different from a clean one in case the artefact spans only over a short interval and thus contributes little to the calculated power. If we compute power on a finer time scale, then data points belonging to an artefact will stand out compared with clean areas. Similarly, human scorers often mark artefacts that span much less than the length of an epoch. Note that the used algorithms do not rely on the length of the scoring epoch. Most algorithms detect artefacts on a sample basis. If a single outlier was detected, then the whole epoch was marked as an artefact. Some work on scoring epochs, but the derived values are independent of the specific epoch length. Thus, the length of an epoch is not relevant for the artefact detection. This also indicates that these algorithms can be applied on a finer time scale than 20 s or 30 s.

We focused on methods that can be applied to single EEG derivations and do not need prior scoring of sleep stages and performed artefact exclusion. They should easily be applicable to large datasets (Luca et al., 2015) as needed to address questions in genetics, epidemiology or precision medicine. Applying one of these algorithms will tremendously reduce analysis time compared with a standard approach (manual scoring). EEGLAB (Delorme & Makeig, 2004; Winkler et al., 2011) provides a large palette of tools to remove artefacts like eye blinks or ECG contamination, but is based on multi-channel (>30) EEG recordings.

We used different measures to assess the performance of the algorithms, among them sensitivity and specificity, area under the ROC curve and average power density spectra. For our applications we selected as optimal parameters those that corresponded to a FPR of approximately 0.1. Even when FPR was set at 0.1, the observed values varied considerably, showing the expected values on average only (Table 2). It should be noted, however, that optimizing one performance measure does not imply that all the other ones are optimized simultaneously. Thus, one needs to decide which aspect has to be optimized.

Although we demonstrated that automatic artefact detection and exclusion works well, in the first place one should aim at obtaining high-quality EEG recordings avoiding as many artefacts as possible.

4.1 | Conclusion

This study demonstrated that simple algorithms work well to automatically detect artefacts in EEG recordings in healthy participants and patients reaching good sensitivity and specificity. They are easily applicable to large datasets, and will speed up data processing tremendously. Many of them even work for on-line data processing and might thus be useful in applications like "closed loop" stimulation during sleep (Fattinger et al., 2017; Ngo, Martinetz, Born, & Molle, 2013).

ACKNOWLEDGEMENTS

This study was supported by a grant of nano-tera.ch (20NA21_145929), of the Swiss National Science Foundation (32003B_146643), the ETH Zurich Research Grant ETHIIRA (ETH-18 11-1), and the NCCR Transfer Projects of the Swiss National Science Foundation (51NF40-1444639).

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

AUTHOR CONTRIBUTIONS

AM and PA designed the analyses; AM conducted the analyses; XO, RR, PA, AW, AW and WJ collected the data; AM and PA wrote the manuscript, and all authors commented and accepted the final version.

ORCID

Peter Achermann http://orcid.org/0000-0002-0208-3511

REFERENCES

Anderer, P., Roberts, S., Schlögl, A., Gruber, G., Klösch, G., Herrmann, W., … Saletu, B. (1999). Artifact processing in computerized analysis of sleep EEG – a review. *Neuropsychobiology*, *40*, 150–157.

Barlow, J. S. (1983). Muscle spike artifact minimization in EEGs by time-domain filtering. *Electroencephalography and Clinical Neurophysiology*, *55*, 487–491.

Barlow, J. S. (1984). EMG artifact minimization during clinical EEG recordings by special analog filtering. *Electroencephalography and Clinical Neurophysiology*, *58*, 161–174.

Barlow, J. S. (1986). Automatic elimination of electrode-pop artifacts in EEGs. *IEEE Transactions on Biomedical Engineering*, *33*, 517–521.

Bodenstein, G., & Praetorius, H. M. (1977). Feature extraction from the electroencephalogram by adaptive segmentation. *Proceedings of the IEEE*, *65*, 642–652.

Comon, P. (1994). Independent component analysis, a new concept? *Signal Processing*, *36*, 287–314.

Coppieters't Wallant, D., Muto, V., Gaggioni, G., Jaspar, M., Chellappa, S. L., Meyer, C., … Phillips, C. (2016). Automatic artifacts and arousals detection in whole-night sleep EEG recordings. *Journal of Neuroscience Methods*, *258*, 124–133.

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*, 9–21.

D'Rozario, A. L., Dungan Ii, G. C., Banks, S., Liu, P. Y., Wong, K. K., Killick, R., … Kim, J. W. (2015). An automated algorithm to identify and reject artefacts for quantitative EEG analysis during sleep in patients with sleep-disordered breathing. *Sleep and Breathing*, *19*, 607–615.

Durka, P. J., Klekowicz, H., Blinowska, K. J., Szelenberger, W., & Niemcewicz, S. (2003). A simple system for detection of EEG artifacts in polysomnographic recordings. *IEEE Transactions on Biomedical Engineering*, *50*, 526–528.

Fattinger, S., De Beukelaar, T. T., Ruddy, K. L., Volk, C., Heyse, N. C., Herbst, J. A., … Huber, R. (2017). Deep sleep maintains learning efficiency of the human brain. *Nature Communications*, *8*, 15405.

Gavelin, R., Klomp, H., Priddle, C., & Uddenfeldt, M. (2004). *Blind source separation – report for adaptive signal processing project*. Sweden: Department of Engineering Sciences, Uppsala University.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, *4*, 1–58.

Gevins, A. S., Yeager, C. L., Diamond, S. L., Spire, J., Zeitlin, G. M., & Gevins, A. H. (1975). Automated analysis of the electrical activity of the human brain (EEG): A progress report. *Proceedings of the IEEE*, *63*, 1382–1399.

Girolami, M. (1998). An alternative perspective on adaptive independent component analysis algorithms. *Neural Computation*, *10*, 2103–2114.

Goncharova, I. I., Mcfarland, D. J., Vaughan, T. M., & Wolpaw, J. R. (2003). EMG contamination of EEG: Spectral and topographical characteristics. *Clinical Neurophysiology*, *114*, 1580–1593.

Gotman, J., Ives, J. R., & Gloor, P. (1981). Frequency content of EEG and EMG at seizure onset: Possibility of removal of EMG artefact by digital filtering. *Electroencephalography and Clinical Neurophysiology*, *52*, 626–639.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: John Wiley & Sons.

Groppe, D. M., Makeig, S., & Kutas, M. (2009). Identifying reliable independent components via split-half comparisons. *NeuroImage*, *45*, 1199–1211.

Habibzadeh, F., Habibzadeh, P., & Yadollahie, M. (2016). On determining the most appropriate test cut-off value: The case of tests with continuous results. *Biochemia Medica*, *26*, 297–307.

Iber, C., Ancoli-Israel, S., Chesson, A., & Quan, S. F. (2007). *The AASM manual for the scoring of sleep and associated events: Rules, terminology and technical specifications*. Westchester, IL: American Academy of Sleep Medicine.

Ktonas, P. Y., Osorio, P. L., & Everett, R. L. (1979). Automated detection of EEG artifacts during sleep: Preprocessing for all-night spectral analysis. *Electroencephalography and Clinical Neurophysiology*, *46*, 382–388.

Lee, T.-W., Girolami, M., & Sejnowski, T. J. (1999). Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources. *Neural Computation*, *11*, 417–441.

Luca, G., Haba Rubio, J., Andries, D., Tobback, N., Vollenweider, P., Waeber, G., … Tafti, M. (2015). Age and gender variations of sleep in subjects without sleep disorders. *Annals of Medicine*, *47*, 482–491.

Ngo, H. V., Martinetz, T., Born, J., & Molle, M. (2013). Auditory closed-loop stimulation of the sleep slow oscillation enhances memory. *Neuron*, *78*, 545–553.

Omlin, X., Crivelli, F., Heinicke, L., Skorucak, J., Malafeev, A., Guerrero, A. F., … Achermann, P. (2018). The effect of a slowly rocking bed on sleep. *Scientific Reports*, *8*, 2156.

Rechtschaffen, A., & Kales, A. (1968). *A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects*. Bethesda, Maryland: National Institutes of Health.

Semlitsch, H. V., Anderer, P., Schuster, P., & Presslich, O. (1986). A solution for reliable and valid reduction of ocular artifacts, applied to the P300 ERP. *Psychophysiology*, *23*, 695–703.

Winkler, I., Haufe, S., & Tangermann, M. (2011). Automatic classification of artifactual ICA-components for artifact removal in EEG signals. *Behavioral and Brain Functions*, *7*, 30.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Malafeev A, Omlin X, Wierzbicka A, et al. Automatic artefact detection in single-channel sleep EEG recordings. *J Sleep Res*. 2018;e12679. https://doi.org/10.1111/jsr.12679