Automatic detection of microsleep episodes with feature-based machine learning

Jelena Skorucak[1,2,3*], Anneke Hertig-Godeschalk[4,5*], David R. Schreier[4,5,6], Alexander Malafeev[1,2], Johannes Mathis[4#], Peter Achermann[1,2,3,7#]

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

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

[3]Sleep and Health Zurich, University of Zurich, Zurich, Switzerland

[4]Department of Neurology, Inselspital, Bern University Hospital, University of Bern, Bern, Switzerland

[5]Graduate School for Health Sciences, University of Bern, Bern, Switzerland

[6]Department of Medicine, Spital STS AG Thun, Switzerland

[7]The KEY Institute for Brain- Mind Research, Department of Psychiatry, Psychotherapy and Psychosomatics, University Hospital of Psychiatry, Zurich, Switzerland

* authors contributed equally

shared last authorship

Corresponding author:

Prof. Dr. Peter Achermann

Institute of Pharmacology and Toxicology, University of Zurich

Winterthurerstrasse 190, 8057 Zurich, Switzerland

peter.achermann@uzh.ch

Abstract

Study Objectives: Microsleep episodes (MSEs) are brief episodes of sleep, mostly defined to be shorter than 15 s. In the electroencephalogram (EEG), MSEs are mainly characterized by a slowing in frequency. The identification of early signs of sleepiness and sleep (e.g. MSEs) is of considerable clinical and practical relevance. Under laboratory conditions, the maintenance of wakefulness test (MWT) is often used for assessing vigilance.

Methods: We analyzed MWT recordings of 76 patients referred to the Sleep-Wake-Epilepsy-Center. MSEs were scored by experts defined by the occurrence of theta dominance on ≥1 occipital derivation lasting 1–15 s, while the eyes were at least 80% closed. We calculated spectrograms using an autoregressive model of order 16 of 1-s epochs moved in 200-ms steps in order to visualize oscillatory activity and derived seven features per derivation: power in delta, theta, alpha and beta bands, ratio theta/(alpha+beta), quantified eye movements, and median frequency. Three algorithms were used for MSE classification: support vector machine (SVM), random forest (RF), and an artificial neural network (long short-term memory [LSTM] network). Data of 53 patients were used for the training of the classifiers, and 23 for testing.

Results: MSEs were identified with a high performance (sensitivity, specificity, precision, accuracy, and Cohen's kappa coefficient). Training revealed that delta power and the ratio theta/(alpha+beta) were most relevant features for the RF classifier and eye movements for the LSTM network.

Conclusions: The automatic detection of MSEs was successful for our EEG-based definition of MSEs, with good performance of all algorithms applied.

KEYWORDS: microsleep; excessive daytime sleepiness; vigilance assessment, maintenance of wakefulness test, machine learning

Statement of Significance

The identification of early signs of sleepiness and sleep is of considerable clinical and practical relevance. We developed methods for the automatic classification of microsleep episodes in a clinical setting using expert scoring and features derived from the electroencephalography and electrooculography. We would like to propose these methods for clinical use as a semi-automatic procedure where automatic scoring would still need to be reviewed and, if necessary, modified by clinical experts. This would lead to a much faster and standardized detection of microsleep episodes.

Introduction

Up to 15-20 % of individuals in the general population suffer from excessive daytime sleepiness (EDS),[1-4] leading to reduced performance at work, and while driving. The main causes for EDS are socially induced sleep deprivation in healthy individuals, medical disorders such as sleep apnea or narcolepsy, and sedative drugs.[5-7] The objective assessment of sleepiness is of high relevance for diagnosis, treatment, and the judgment of fitness to drive. Even though sleep-wake medicine profited from the recent technological progress, the objective assessment of sleepiness still remains a challenge.

Up to now, the gold standard to objectively assess sleep and wakefulness is based on polysomnographic (PSG) data, in particular on the electroencephalogram (EEG). Visual sleep scoring criteria were initially established by Rechtschaffen and Kales [8] in 1963, and are currently applied in a version which was adapted and amended by the American Academy of Sleep Medicine (AASM).[9,10] These criteria are based on 30-s epochs, which are classified into wakefulness, rapid eye movement (REM) sleep, and non-rapid eye movement (NREM) sleep stages 1-3 (N1-N3). The multiple sleep latency test (MSLT [11]) and the maintenance of wakefulness test (MWT [12]) are clinically applied to assess EDS.[5,13] The MWT evaluates the patient's ability to resist falling asleep despite the presence of EDS and is considered to be the most important

vigilance test to assess patient's fitness to drive.[12,14,15] Whether it is still accurate to classify wakefulness and sleep based on 30-s epochs is debatable [16], especially in the context of driving where short lapses can have fatal consequences. Thus, the term "microsleep" appears more often in today's scientific literature and mostly refers to "sleep" of < 15 s duration derived from PSG data, but microsleep can also be based on behavior assessed by videography, such as eye lid closure, or based on psychomotor performance tests. EEG derived microsleep episodes (MSEs) are visually scored as 3 to 15 s periods dominated by theta activity (EEG power in the 4 – 8 Hz frequency range) that replaces alpha activity (power in the 8 – 12 Hz range) and often accompanied by eye lid closure. Also less precise definitions were used for MSEs, such as "short-lasting burst of typical stage 1 sleep".[17-21] Only rarely MSEs shorter than one second are taken into account.[22] Besides the lack of standardization and the different approaches used for MSE identification, visual scoring is time consuming, requires training and experience, and remains subjective.

State of the art of the automatic vigilance detection

Algorithms have been developed to track vigilance and to detect MSEs based on electrophysiological (mainly EEG) and videography data. Already in 1997, automatic estimation of alertness levels during an auditory monitoring task was

performed using EEG data and a neural network approach with spectral data as input.[23] Among the increasing number of EEG-based algorithms developed, the Vigilance Algorithm Leipzig (VIGALL) has become popular to track vigilance (i.e. vigilance regulation) in health and disease.[24,25] After artefact removal, 1-s EEG segments of multiple channels were classified into 7 different stages of vigilance reaching from fully awake to sleep. Sleep latency in the MSLT was correlated with the vigilance score predicted by the VIGALL measured in the wake EEG recording after the MSLT, showing a moderate correlation between these two measures.[24] Other studies performed drowsiness detection based on a single EEG channel with an artificial neural network approach and spectral or wavelet derived features,[26,27] or based on a means comparison test to detect changes of relative power in different frequency bands.[28,29] The classification of drowsiness in these studies was performed on 1-s,[29] 5-s,[26] or 10-s[28] epochs, and expert scoring was performed on 20, 30, and 30 s, respectively. Sauvet et al. [28] detected MSEs in pilots during long-haul overnight flights. Classification was performed to discriminate "awake" and "sleepy", where "sleepy" was defined as any sleep stage (N1-N3, or REM sleep) and a sensitivity of 87 %, an accuracy of 98 %, and a kappa of 0.94 was reported. The aim of some other studies was to discriminate wakefulness and N1 in PSG data recorded during the night.[26,27,30] Garces Correa et al. [26] obtained an average of 85.5 % of correct detections with a neural network approach using spectral analysis features. Sriraam et al. [30] used

spectral entropy and a multilayer perceptron feed forward neural network and reported an accuracy of 99.2 %. Belakhdar et al. [27] achieved an accuracy of approximately 89 % using a multi-layer perceptron and spectral power in 1-Hz bands. However, detecting N1, which is usually scored in 20-s or 30-s epochs, may differ from detecting short MSEs.

Other algorithms used the EEG in combination with videography and performance testing to detect MSEs. For example, Peiris et al. [31] estimated the fractal dimension of the EEG to detect behavioral MSEs, which were identified by experts based on face videography and lapses in a tracking task, with a weak correlation between automatic detection and expert scoring. Further algorithms detected behavioral performance lapses on a second resolution based on spectral EEG features, as well as facial video recordings and tracking task performance.[32,33] Davidson et al. [33] used long short–term memory (LSTM) recurrent neural network reaching a sensitivity of 48 % and specificity of 93 %, while Peiris et al. [32] used linear discriminant analysis, and obtained a sensitivity of 73.5 % and a specificity of 25.5 %. These two studies applied interesting approaches but did not obtain a good performance, and detection was based on task performance and thus, were not suited as benchmark for our study. Golz and colleagues [34] used EEG data recorded in a driving simulator for MSE classification (detection and prediction) based on support vector machines and optimized learning vector quantization. Expert scoring of MSEs was based on visual

inspections of video material, of lane deviation time series and of the electrooculogram (EOG). Input data were features derived from spectral power or the Choi-Williams distribution of 8-s EEG segments reaching accuracies >80 %. Another study developed online detection of MSEs,[29] based on a means comparison test to detect changes in relative alpha power, with a sensitivity of 85 % and specificity of 80 %.

To the best of our knowledge, automatic detection of MSEs was not investigated in a clinical setting with the commonly used MWT. The aim of this study was to develop machine learning based algorithms to automatically detect MSEs in a clinical setting using features derived from EEG and EOG data.

Methods

Patients

Seventy-six patients that were suspected to have excessive daytime sleepiness (EDS) and consequently underwent a MWT were analyzed. They were randomly selected out of patients who had been referred to the Sleep-Wake-Epilepsy-Centre, Bern University Hospital, Inselspital. Patients with a large diversity of suspected diagnoses were included: excessive daytime sleepiness, sleep apnea, narcolepsy, idiopathic hypersomnia, non-organic hypersomnia, insomnia, and others (Table 1).

No medical in- or exclusion criteria for the patients existed. Subgroups of patients were not selected for this study since only few patients were available with a certain suspected diagnosis due to their low prevalence (e.g. prevalence for narcolepsy is 25-50 per 100,000 people),[35,36] and the algorithms should be valid independent of any disorder or medication. Further, variability in the EEG recorded during the MWT is mainly related to the severity of sleepiness.[37] The mean age of the patients was 45.6 years (range: 18.0 – 81.4 years), and 50 of them were male, 97 % were Caucasian, and approximately 1/3 were obese, mostly sleep apnea patients.

The study was conducted according to the Declaration of Helsinki, Swiss Law, and the ethical approval of the local ethics committee (KEK-Nr. 308/15). Data were included based on a general consent that patients signed with the hospital.

Assessment

As part of the clinical routine procedures, patients underwent four 40-min MWT trials in one day (starting at approximately 8:00, 10:00, 15:00 and 17:00). Since visual scoring was very time consuming (see below), only the MWT recorded at 15:00 (MWT-3) was analyzed. MSEs were most likely to occur in this trial according to clinical experience, and might be related to a circadian or time of day contribution (mid-afternoon or post-lunch dip). In the MWT, patients were seated on a chair in a semi-darkened room (0.1 Lux at corneal level) and were instructed to stay awake for

as long as possible without any interaction or activities. Each trial lasted 40 min, and it was supposed to be terminated earlier if three consecutive 30-s epochs of N1 or one epoch of any other sleep stage was observed. However, if the laboratory technician missed to terminate the recording due to the appearance of sleep epochs, data from the entire recording were used in this study for training and testing of the classifiers in order to obtain as much data as possible (i.e. also including sleep episodes longer than 15 s).

EEG recordings and data pre-processing

A standard EEG, EOG, submental electromyography (EMG), electrocardiography (EKG, 2 electrodes placed subclavicular (right) and on the lateral thorax on the approximate height of the heart point (left)), respiratory flow, and face videography including audio were simultaneously recorded. EEG electrodes were placed according to the 10-20 electrode placement system,[38] at sites O1-M2, O2-M1, C3-M2, C4-M1, CZ-M1, F3-M2, F4-M1 (referenced to the contralateral mastoids). Impedance values were at or below 5 kΩ at the beginning of the recordings.

Data were recorded using RemLogic™ (Embla Systems LLC) devices. The sampling and storage rates were 200 Hz, and the following hardware filters were applied: a

high-pass at 0.3 Hz, a low-pass at 70 Hz, and a notch filter at 50 Hz. Data were exported in the European Data Format (EDF) for further processing.

EKG artefacts contaminating the EEG were removed using a procedure modified from Purcell et al. [39]: first the EKG pattern in the EEG was calculated (moving window; triggered with the R peak of the EKG), and next the corresponding pattern was subtracted from the EEG (see Supplementary methods). This procedure was applied to all recordings irrespective of whether EKG artifacts were clearly visible in the EEG or not.

The quantitative analysis was performed in MATLAB R2018a (The Math Works Inc., Natick, MA, USA), using left and right occipital EEG derivations (O1-M2 and O2-M1) and left and right EOG derivations. We focused in a first step on occipital channels as the alpha rhythm, present during rest with eyes closed, originates from the occipital lobes of the brain. Further, the wake-sleep transition zone characterized by the loss of alpha activity and shift to theta activity is best seen in the occipital channels.[40]

Visual scoring

The scoring was conducted by an experienced scorer (see Hertig-Godeschalk et al.[40] for details) and in in around 2/3 of the trials, the final scoring was verified by other

experienced scorers and differences were resolved by discussions. MSEs (visible in both channels), unilateral MSEs, microsleep episode candidates (MSEc) or episodes of drowsiness (ED) were scored as defined in Bern continuous and high-resolution wake-sleep (BERN) scoring criteria.[40] MSEs were scored based on occipital EEG derivations (O1-M2, O2-M1), EOG, and videography. MSEs were visually defined as episodes of 1-15 s duration with clear slowing in the EEG with a theta dominance similar to N1, and eyes at least 80 % closed (visually determined from face videography). MSEs were typically preceded by slow eye movements in the EOG. If a MSE fulfilled all criteria only at one occipital channel, it was categorized as a unilateral MSE. Borderline EEG sections between clear wakefulness and MSEs were categorized as MSEc or as ED that were particularly difficult to score (see [40]). This time-consuming visual scoring resulted in a total of 1262 MSEs and segments of sleep.

Power spectral analysis

Spectral analysis of the EEG was performed using an autoregressive model of order 16 (Burg method [41]). A 1-s sliding window was moved through the data in steps of 200 ms. This approach allows high temporal resolution and good visualization of oscillatory activity such as alpha or theta activity.[42] The model order was chosen

based on our experience. For the automatic detection of oscillatory events order 8 was applied.[43,44] However, to illustrate oscillations in the spectrogram we experienced that order 16 is better suited (see Figure 1 in Olbrich et al.[45]). Figure 1 illustrates 20 s of an EEG signal (O2-M1, upper panel), and the corresponding spectrogram (lower panel) with a MSE occurring between the two vertical red lines. Oscillatory alpha activity (10 Hz) is clearly visible.

Feature engineering

Feature engineering is the process of extracting quantifiable properties from the data that will serve as an input for the classification algorithms. Furthermore, features may serve as objective markers to support scoring of MSEs.[40] Although there is still an ongoing discussion about the best markers and criteria for MSE detection, most of the studies agree that the alpha and theta bands, as well as slowing of eye movements (i.e. rolling eye movements) and a lack of eye blinks are good indicators.[29,46] The disappearance of alpha activity in the EEG is predominantly seen in the posterior region of the brain.[29] In our study, different MSE markers were identified from two occipital EEG derivations, and from the EOG. The occurrence of eye movements was quantified by the ratio of delta power of the EOG (difference between two EOG channels) and delta power of the EEG from occipital derivations.[47] This is a rough and simple overall quantification of eye movements that does not

allow to dissociate different kind of eye movements or lid blinks. Measures derived from the EEG were: power in the delta (0.8 – 4 Hz), theta (4 – 8 Hz), alpha (8 – 12 Hz) and beta (12 – 26 Hz) bands, the ratio theta/(alpha+beta) (T/AB), and the median frequency in the 0.8 – 26 Hz range (Figure 2). Power in the delta, alpha, theta and beta were smoothed by a 1-s moving median filter. These features proved to be helpful for the visual scoring of MSEs.[40] The seven features mentioned above were calculated from left and right occipital EEG derivations (O1-M2 and O2-M1). Features of both derivations were used as input for classification algorithms, resulting in a total of 14 features sampled every 200 ms (see spectral analysis).

Training of the classifiers, testing, and post-processing

We applied three classifiers: a long short-term memory recurrent neural network (LSTM [48,49]), random forest (RF, 100 trees [50]), and a support vector machine (SVM, radial basis kernel [51]). Recurrent neural networks are taking the temporal structure into account and therefore have a good performance for time series data.[52] LSTMs as well as other artificial neural networks usually consist of an input layer (having the size of the feature vector), one or more hidden layers, and the output layer. The structure of our LSTM was as follows: an input layer (14 neurons), 2 LSTM layers (100 neurons each) each followed by a dropout layer (dropout probability 0.3),

followed by a fully connected layer (2 neurons), a softmax layer, and classification output layer. Sixteen training epochs were applied, i.e. the entire training data was passed through the neural network 16 times. The adaptive moment estimation optimization algorithm (Adam) was used to update network weights during training.[53] The input of the LSTM consisted of a moving time window of 9 s (45 samples; step 200 ms).

The number of trees of the RF, SVM kernel functions and LSTM architecture were optimized (manual tuning) on a smaller data set (when not all data were scored yet), and 100 trees, the best kernel and model were applied to the final data set.

According to Oshiro et al.[54] 64 to 128 trees would be sufficient for medical data without any performance gain with a further increase in the number of trees. We tested with 50 and 100 trees and did not observe a substantial performance increase. Thus, we finally selected 100 trees. For the SVM we compared the linear and radial basis kernel with a better performance for the radial basis kernel. For the LSTM network the number of neurons in the hidden layer, the number of hidden layers, the inclusion and probability of the dropout layers, a bidirectional/unidirectional architecture, and the window size were tested. Best performance was achieved with the above-mentioned architecture and parameters.

The classifiers were trained on 53 patients (70 %; 18 without MSEs) and tested on 23 patients (30 %; 12 without MSEs). Patients were randomly assigned to the training

and testing sets, and data of a patient contributed to only one set (either training or testing). Only bilateral MSEs and wakefulness were included for training (unilateral MSEs, MSEc and ED were excluded). All data contributed to the training of the LSTM, while data were balanced for the training of the RF and the SVM such that the same number of 200-ms data points of MSE and wakefulness categories were used, i.e. all data points corresponding to MSEs were included and the same number of data points were randomly selected from wakefulness data. Balancing was performed across the pooled data since some patients did not have MSEs. After classification (at 200-ms steps), identified MSEs shorter than 1 s were excluded before comparison with visual scoring. Furthermore, we applied smoothing with a 9-s moving median filter to the SVM and RF classifications in order to account for the temporal structure of the data. This time interval was selected to be the same as the one used in the LSTM neural network.

Assessment of classification performance

Performance of the classifiers was assessed by determining specificity, sensitivity, precision, accuracy, and the Cohen's kappa coefficient.[55-59] The human scoring was converted to same temporal resolution (200 ms) of the features. Sensitivity represents the true positive rate (i.e. the proportion of MSEs that are correctly identified – true

positives divided by the sum of the true positives and false negatives), and specificity stands for the true negative rate (i.e. the proportion of wakefulness that are correctly identified – true negatives divided by the sum of the true negatives and false positives). Accuracy is a measure combining sensitivity and specificity (correctly identified positives and negatives divided by the sum of the correctly and incorrectly identified ones). Specificity and accuracy are biased measures and are only reported for comparison with published data. Precision is a ratio of true positives and the combination of true and false positives. The Cohen's kappa coefficient is a more robust measure than accuracy, which takes the possibility of the agreement occurring by chance into account.[59] Interpretation of the performance results for Cohen's kappa was made using Landis and Koch levels [60]: <0.00 – poor; 0.00-0.20 – slight; 0.21-0.40 – fair; 0.41-0.60 – moderate; 0.61-0.80 – substantial; 0.81-1.00 – almost perfect identification.

Training of the classifiers was performed by taking only bilateral MSEs and wakefulness into account. Testing was performed on the entire MWT-3, and performance was estimated based on only bilateral MSEs and wakefulness, or considering unilateral MSEs, MSEc and ED either as wakefulness or MSEs (see Table 2 and S1 for the different combinations applied). Overall performance measures across all patients (pooled data) and mean values across patients are reported. Recordings not having MSEs had to be excluded for the calculation of mean

sensitivity, precision and Cohen's kappa since these measures take into account positives (i.e. MSEs). Individual performance measures are reported in supplementary Figure S3.

Assessment of inter-scorer variability

Out of 23 patient recordings used for testing performance of the algorithms, 5 were scored independently by 2 different scorers. These records were randomly selected from those recordings in the test data set which had MSEs. Performance measures were calculated in the same way as for the algorithms.

Assessment of the importance of the features

During training the RF classifier constructs a variety of decision trees. It is possible to rank the importance of features contributing to the classification with RF, which can bring new knowledge about the properties of the data. The RF uses a "tree bagging" algorithm,[50] which takes a random subset of data from a training set, and creates a decision tree for each random subset. In order to create the decision trees, the RF selects a random subset of features at each node of the tree (decision split). Feature importance was calculated as the increase in prediction error if the values of

the corresponding feature were permuted. This measure was computed for every tree, then averaged over the entire set of decision trees and divided by the standard deviation over the entire set of decision trees (TreeBagger class, Matlab R2018a). Feature importance was also assessed for the training of the selected LSTM network. Feature permutation was performed (one feature at a time) in the training set and 7 models were trained, each with one of the features "destroyed" and we determined the corresponding accuracy and loss functions, and model performance (overall Cohen's kappa).

In addition to the feature importance during training, we determined how corrupted features affect classification of the test set. Thus, we performed feature permutation ("destroyed" features) one at a time in the test set and calculated performance with Cohen's kappa for the three algorithms (LSTM, RF and SVM).

Results

One example of a MSE in the EEG with the corresponding spectrogram is plotted in Figure 1. Beginning and end of a MSE are marked with vertical red lines. Alpha activity was present just before the MSE and thereafter, but not during the MSE. Alpha activity is evident in spectrogram as high power at around 10 Hz (red color; Figure 1). Furthermore, there was a drop in beta activity during the MSE, visible in

spectrogram as low power above 12 Hz (dark blue areas). Moreover, appearance of theta activity is evident as high power in 4 – 8 Hz range (yellow color) during the MSE.

The different features, mostly derived from power spectra, serving as input vectors for the classifiers are depicted in Figure 2 (3 min exemplary data of one patient). Alpha and beta activity decreased during MSEs. Although theta activity was not clearly increased, the ratio theta/(alpha+beta) revealed an increase during MSEs. The median frequency indicated the slowing of the EEG during MSEs, and eye movements were mostly lacking.

Classification performance

MSE detection in one patient with the three classifiers and the corresponding expert scoring are illustrated in Figure 3. Only bilateral MSEs and wakefulness are plotted (excluding unilateral MSEs, MSEc and ED; i.e. time axis is compressed). The entire recording with all the scored categories is provided in supplementary Figure S3 (Patient 22).

All three feature-based classifiers showed good performance (high sensitivity, specificity, precision, accuracy, Cohen's kappa) with e.g. kappa coefficients ranging from 0.75 to 0.83 when considering only bilateral MSEs and wakefulness (Table

1A), i.e. reflecting substantial to almost perfect identification.[60] The mean duration of false positives (MSEs) amounted to 1.10 ± 0.29 (SD) min (n= 23).

For exploratory purposes, we also calculated performance considering not only bilateral MSEs, but also assigning unilateral MSEs, MSE candidates, and ED to the category MSE or to wakefulness when calculating performance (Table 1B, 1C). This reduced the performance estimates, i.e. kappa values became moderate. Assigning MSEc to the category MSE and ED to the category wakefulness resulted in a substantial performance (Table 1D) indicating that MSEc might be closer to MSE and ED closer to wakefulness.

Detection of MSEs in the entire MWT-3 of all 23 patients in the test set are illustrated in Supplementary Figure S3, with the expert scored categories at the top (red). Some of the false positive MSEs coincided with MSEc or ED. To quantify this correspondence, MSE detection performance was evaluated either against MSEc (Table S1A) or ED (Table S1B). The low performance indicates that detected MSEs only partially correspond to MSEc or ED.

Our three algorithms had a high ability to correctly identify MSEs. Overall, all three classifiers performed well, although the LSTM showed generally a better performance than the SVM and RF classifiers (Tables 1, S1).

Importance of the different features

During training, the RF classifier provides information on the importance of different features for the classification (Figure 4A). The ratio theta/(alpha+beta) (increase), delta activity (increase), and beta activity (reduction) had the highest contributions. This was expected from the expert scoring criteria where experts score MSEs according to the slowing of the EEG, a shift from the alpha to the theta range.[40] However, by visual inspection of Figure 2, one might have expected that the slowing of the EEG (median frequency) is also an important feature. Destroying features in the training of the LSTM neural network resulted in a hardly affected Cohen's kappa (Figure 4B) with eye movement showing a small decrease of kappa, and accuracy and loss functions were very similar (Figure S3), indicating that neural networks are quite robust.

Destroying features in the test set (i.e. corresponding to corrupted features) revealed that eye movements were of importance for all three algorithms (Figure 4 C, D, E). In addition, the ratio theta/(alpha+beta) was important for the RF algorithm (Figure 4D) and alpha activity for the SVM (Figure 4E).

Discussion

The three algorithms developed for the automatic detection of MSEs showed a good performance, indicating that reliable computerised MSE detection is feasible based only on EEG and EOG data. To our knowledge, this is the first study to automatically detect MSEs in a clinical setting (MWT) in which visual scoring of MSEs is routinely performed. The concept behind the definition of sleep-like episodes (bilateral and unilateral MSEs, MSEc and ED) representing different levels of sleepiness is presented in Hertig-Godeschalk et al.[40]

Automatic classification slightly outperformed human scoring in performance, having an average Cohen's kappa coefficient of 0.68 (LSTM), while the human inter-scorer kappa was 0.67 (average of 5 recordings). MSEs are short fragments (1–15 s) of sleep stage N1 scored in 30-s epochs. Interestingly, the inter-scorer agreement for MSEs was higher than the one reported for N1: Cohen's kappa coefficient for the Rechtschaffen and Kales scoring was 0.35 [61] and for the AASM scoring 0.31 [62] or around 0.60.[63,64] Further, automatic scoring of N1[65,66] was worse than the performance of our algorithms. This indicates that our visual MSE scoring was precise.

In this study, the analysis was performed on two occipital EEG derivations and a bipolar EOG derivation. The occipital region was selected as region of interest since

clinical scoring is also performed on occipital channels and features of the MSEs are often best visible in this brain region. Nevertheless, considering local aspects of sleep, in future applications it may be of importance to apply the algorithm to other brain regions. Furthermore, including further brain regions might lead to a better discrimination between wakefulness and EDs or MSEc.

The features used as input for classifiers were mostly EEG power in different frequency bands (e.g. delta, theta, alpha, beta), which are well-established and commonly used features for MSE or drowsiness classification.[32,33,67-69] Besides these well-established features we also defined the ratio theta/(alpha+beta) that was of importance for the RF classification. In addition, eye movements were quantified and median frequency of the EEG between 0.8 – 26 Hz was calculated to track the slowing of the EEG frequency during MSEs. These features were selected based on expert experience, literature, and from inspecting numerous spectrograms.[40] Corrupted eye movements reduced the quality of classification of all three algorithms. We are also working on detecting MSEs based on raw EEG/EOG data with deep learning, i.e. features are "learned" by the artificial neuronal network applied.[65]

All classifiers showed a good performance (Table 1). The classifier with the best performance was the LSTM neural network, with an average Cohen's kappa coefficient of 0.83 (only MSEs and wakefulness; almost perfect identification) or

0.68 (unilateral MSEs, MSEc and ED assigned to wakefulness; substantial identification). In contrast to the SVM and RF, LSTM neural networks take the temporal context into account. The LSTM network had a 9-s memory, while SVM and RF classified single 200-ms intervals independently of each other (picking up information of 1 s), which were afterwards smoothed with a 9-s moving median filter.

Features were sampled at 200-ms intervals. However, as a 1-s window was used for spectral analyses and moved through the data, detected events are always at least 1 s long and their beginning and end is smeared. Additionally, the automatic classification uses a binary system (MSE, wakefulness) while scorers are confronted with gradual changes which sometimes make it hard to clearly define the beginning and the end of a MSE. Detected MSEs that start or end a bit earlier or later than the scored ones will lead to a penalty in the performance, although this is not clinically relevant. Furthermore, features of the episodes scored as unilateral MSEs, MSEc, or ED may have very similar features to MSEs or wakefulness, depending on the case. In order to reduce above-mentioned problem (avoid ambiguity), only data scored as MSEs and wakefulness were used for the training of the classifiers.

MSEs (positives) are rare compared to wakefulness (negatives). Therefore, measures like specificity and accuracy that consider correctly identified negatives are highly biased (e.g. if the classification algorithm predicts only wakefulness, it will have high

sensitivity and accuracy due to the correct prediction of the majority of the data points, which will not be informative about how well MSEs were detected). Therefore, it is important to calculate measures taking positives into account, like sensitivity and precision. Precision informs about the appearance of false positives in the prediction. It differs from sensitivity which informs about how many MSEs are correctly identified compared to all scored MSEs. We consider sensitivity, precision and the Cohen's kappa coefficient as relevant measures for our application, but we still report specificity and accuracy since these two measures were often reported in the literature regarding MSE or drowsiness detection.[26,27,29,30,32-34] Further, performance measures carry different information, and it is important to assess them together to get the most optimal impression about the performance of the classification. It is essential that the algorithm does not detect a large number of false positives in patients not having any MSEs, therefore, we also report overall performance (pooled data).

Although MSE detection worked generally very well in most of the patients, in three patients we identified a higher amount of false positive MSEs (Figure S3, Patients 14, 16 and 19) with 4.0, 3.3 and 4.4 min of false positives. However, we do not consider false positives to be a systematic problem of the algorithm since the classifier worked well in patients which did not have any MSEs, i.e. hardly any MSEs

were detected. In 11 out of 12 patients not having any MSEs, also no MSEs were detected with at least one of the three classifiers (Figure S3, Patients 1-11).

Inter-patient variability in automatic MSE detection was also reflected in the human scoring, while in some cases humans agree on most of the episodes, in others there was a lot of variability. Therefore, we consider the variability in the algorithm performance to be due to the data itself and not due to the algorithm.

Our classifiers were more sensitive than precise, i.e. they detected most of the MSEs without missing MSEs, but they detected more MSEs than they should, resulting in false positives. For the clinical application of MSE detection, it is more favorable to have higher sensitivity and not to miss MSEs. Furthermore, the clinical scoring of the MSEs was rather conservative, and it could well be that some of the false positive MSEs were real MSEs missed by the human scorer. Moreover, the visually scored MSEs were considered here as a sort of "gold standard" for the training and the validation of the automatic detection of MSEs, being aware of the great uncertainty in the visual scoring (especially if done by a single expert).

Our performance mostly exceeded performance obtained with other algorithms reported in the literature.[26,27,29,32-34] However, MSE detection was performed in different settings across the different studies. Some studies used sleep recordings to discriminate between wakefulness and N1 sleep,[26,27,29,30] while others analyzed data recorded in a driving simulator,[34] or even recordings of real-life situations in pilots

during long-haul overnight flights.[28] Further, MSE scoring in some studies was EEG-based,[26,27,30] and in others based on behavioral lapses and videography.[32-34]

Our algorithms could be used as a semi-automatic procedure in clinical practice and research laboratories, where the automatically detected MSEs could be quickly checked by clinical experts. This would result in a much faster evaluation of the recordings compared to the traditional visual scoring, and in more standardized and replicable results.

Limitations and outlook

The classification algorithms developed in this paper were trained using expert labels. Human scoring of MSEs is very time consuming and there was no capacity for scoring the data by multiple experts in the scope of this study. Therefore, the algorithms were trained using labels only of one expert. Ideally, data would be scored independently by multiple scorers and the consensus scores would be used for training and testing (in a form of probability of being in a specific stage), which would lead to more generalizable algorithms. However, training based on a very experienced scorer may lead to better performance than training based on multiple less experienced scorers which could result in more noise or even decrease the performance of the algorithm.

Furthermore, a feature-based approach was applied, but there are deep learning based approaches that automatically detect optimal features and perform classification at the same time with raw data. Some of these cutting-edge approaches were used for reliable automatic sleep stage scoring,[70,65,66] even with a higher time resolution of 5 s than the classical 30 s and speeded up the diagnosis of type 1 narcolepsy.[66]

Our approach was developed with classical occipital EEG derivations. Whether MSE detection works equally well in other derivations or with around-the-ear EEG recordings[71] needs to be explored in the future.

Further, only one MWT per patient (the one at 15:00) was used in this study due to the limitations of the time-consuming human scoring. The additional MWT recordings could be automatically scored and might reveal time-of-day influences on the occurrence of MSEs.

In this study MSEs were defined mainly based on the EEG, and behavioral lapses and their connection to MSEs were not investigated. It may be of interest to apply these algorithms in a driving simulator setting and compare the detected MSEs with behavioral lapses (e.g. off-road events). Moreover, MSEs detected in an MWT could be related to driving performance in a driving simulator.

Conclusion

We proved that MSEs can reliably be detected with machine learning, applying classical (SVM or RF) as well as state-of-the-art deep learning algorithms (LSTM). RF and SVM classifiers revealed a similar performance, while classification with a LSTM resulted in slightly better performance. Interestingly, this performance was achieved with a mainly EEG centred approach, while the human scorer used face videography in addition. Our algorithms are well suited for a semi-automatic application in a clinical setting, i.e. automatic MSE detection in a first step and next, the validation by experts. This would lead to a much faster and more standardized detection of MSEs as there is currently no agreement in the field about MSE scoring. In most clinical sleep labs, MSEs are not scored due to the ambiguity and the time-consuming procedure. Instead, sleep is scored in 30-s epochs. However, short sleep of 1-3 s (i.e. MSE) may have fatal consequences e.g. while driving. We proposed criteria for MSE scoring[40] on which these algorithms were trained. What we are hoping for is that the proposed scoring criteria and the automatic MSE detection will increase the attention on the wake-sleep transition zone, encourage clinicians to assess MSEs in their daily work, and open new doors for fitness-to-drive assessments.

Acknowledgments and funding

This work was supported by the Swiss National Science Foundation (SNSF, grants 32003B_146643 and 32003B_176323), nano-tera.ch (grant 20NA21_145929), Clinical Research Priority Program "Sleep and Health" of the University of Zurich, and the Swiss Commission of Technology and Innovation (CTI; grant 17864.1 PFLS-LS).

Algorithms and data availability

The LSTM, RF and SVM classifiers are available as Matlab structure (supplementary classifiers.mat file) and instructions on their use are provided in supplementary material. Data are available in the Zenodo repository (DOI: 10.5281/zenodo.3251716).

Disclosure Statement

Financial Disclosure: none.

Non-financial Disclosure: none.

References

1. Ford ES, Cunningham TJ, Giles WH, Croft JB. Trends in insomnia and excessive daytime sleepiness among U.S. adults from 2002 to 2012. Sleep Med. 2015; 16 (3): 372–378.
2. Hayley AC, Williams LJ, Kennedy GA, et al. Excessive daytime sleepiness and falls among older men and women: cross-sectional examination of a population-based sample. BMC Geriatr. 2015; 15: 74.
3. Hara C, Lopes Rocha F, Lima-Costa MFF. Prevalence of excessive daytime sleepiness and associated factors in a Brazilian community: The Bambuí study. Sleep Med. 2004; 5 (1): 31–36.
4. Young TB. Epidemiology of daytime sleepiness: definitions, symptomatology, and prevalence. J Clin Psychiatry. 2004; 65 Suppl 16: 12-16.
5. Mathis J, Hess CW. Sleepiness and vigilance tests. Swiss Med Wkly. 2009; 139 (15-16): 214–219.
6. Akerstedt T, Bassetti C, Cirignotta F, et al. *Sleepiness at the wheel - white paper*. 2013.
7. Gottlieb DJ, Ellenbogen JM, Bianchi MT, Czeisler CA. Sleep deficiency and motor vehicle crash risk in the general population: a prospective cohort study. BMC Med. 2018; 16 (1): 44.
8. Rechtschaffen A, Kales A. A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects. Los Angeles: UCLA Brain information service/Brain research institute. 1963.
9. Berry RB BR, Gamaldo CE, Harding SM, Lloyd RM, Marcus CL, et al. The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications. Version 2.5. Darien, IL: American Academy of Sleep Medicine. 2018.
10. Iber C. *The AASM manual for the scoring of sleep and associated events: Rules, terminology and technical specifications.* Westchester, IL: American Academy of Sleep Medicine; 2007.
11. Carskadon MA, Dement WC, Mitler MM, Roth T, Westbrook PR, Keenan S. Guidelines for the multiple sleep latency test (MSLT): a standard measure of sleepiness. Sleep. 1986; 9 (4): 519–524.
12. Mitler MM, Gujavarty KS, Browman CP. Maintenance of wakefulness test: a polysomnographic technique for evaluation treatment efficacy in patients with excessive somnolence. Electroencephalogr Clin Neurophysiol. 1982; 53 (6): 658–661.
13. Mathis J, de Lacy S, Roth C. Measuring - monitoring sleep and wakefulness. In: Bassetti CL, ed. ESRS European sleep medicine textbook. Regensburg: European sleep research society; 2014: 125–143.

14. Littner MR, Kushida C, Wise M, et al. Practice parameters for clinical use of the multiple sleep latency test and the maintenance of wakefulness test. Sleep. 2005; 28 (1): 113–121.
15. Mathis J, Schreier D. Daytime sleepiness and driving behaviour. Therapeutische Umschau Revue thérapeutique. 2014; 71 (11): 679–686.
16. Harrison Y, Horne JA. Occurrence of "microsleeps' during daytime sleep onset in normal subjects. Electroencephalogr Clin Neurophysiol. 1996; 98 (5): 411–416.
17. Guilleminault C, Billiard M, Montplaisir J, Dement WC. Altered states of consciousness in disorders of daytime sleepiness. J Neurol Sci. 1975; 26 (3): 377–393.
18. Tirunahari VL, Zaidi SA, Sharma R, Skurnick J, Ashtyani H. Microsleep and sleepiness: a comparison of multiple sleep latency test and scoring of microsleep as a diagnostic test for excessive daytime sleepiness. Sleep Med. 2003; 4 (1): 63–67.
19. Moller HJ, Kayumov L, Bulmash EL, Nhan J, Shapiro CM. Simulator performance, microsleep episodes, and subjective sleepiness: normative data using convergent methodologies to assess driver drowsiness. J Psychosom Res. 2006; 61 (3): 335–342.
20. Boyle LN, Tippin J, Paul A, Rizzo M. Driver Performance in the Moments Surrounding a Microsleep. Transportation research. 2008; 11 (2): 126–136.
21. Herrmann US, Hess CW, Guggisberg AG, Roth C, Gugger M, Mathis J. Sleepiness is not always perceived before falling asleep in healthy, sleep-deprived subjects. Sleep Med. 2010; 11 (8): 747–751.
22. Poudel GR, Innes CRH, Bones PJ, Watts R, Jones RD. Losing the struggle to stay awake: Divergent thalamic and cortical activity during microsleeps. Hum Brain Mapp. 2014; 35: 257–269.
23. Jung TP, Makeig S, Stensmo M, Sejnowski TJ. Estimating alertness from the EEG power spectrum. IEEE transactions on bio-medical engineering. 1997; 44 (1): 60–69.
24. Olbrich S, Fischer MM, Sander C, Hegerl U, Wirtz H, Bosse-Henck A. Objective markers for sleep propensity: comparison between the Multiple Sleep Latency Test and the Vigilance Algorithm Leipzig. J Sleep Res. 2015; 24 (4): 450–457.
25. Olbrich S, Sander C, Minkwitz J, et al. EEG vigilance regulation patterns and their discriminative power to separate patients with major depression from healthy controls. Neuropsychobiology. 2012; 65 (4): 188–194.
26. Garces Correa A, Orosco L, Laciar E. Automatic detection of drowsiness in EEG records based on multimodal analysis. Med Eng Phys. 2014; 36 (2): 244–249.
27. Belakhdar I, Kaaniche W, Djemal R, Ouni B. Single-channel-based automatic drowsiness detection architecture with a reduced number of EEG features. Microprocessors and Microsystems. 2018; 58: 13-23.
28. Sauvet F, Bougard C, Coroenne M, et al. In-flight automatic detection of vigilance states using a single EEG channel. IEEE transactions on bio-medical engineering. 2014; 61 (12): 2840–2847.
29. Picot A, Charbonnier S, Caplier A. On-line automatic detection of driver drowsiness using a single electroencephalographic channel. Conference proceedings : Annual

International Conference of the IEEE Engineering in Medicine and Biology Society IEEE Engineering in Medicine and Biology Society Conference. 2008; 2008: 3864–3867.
30. Sriraam N, Padma Shri TK, Maheshwari U. Recognition of wake-sleep stage 1 multichannel eeg patterns using spectral entropy features for drowsiness detection. Australas Phys Eng Sci Med. 2016; 39 (3): 797–806.
31. Peiris MR, Jones RD, Davidson PR, Bones PJ, Myall DJ. Fractal dimension of the EEG for detection of behavioural microsleeps. Conference proceedings : Annual International Conference of the IEEE Engineering in Medicine and Biology Society IEEE Engineering in Medicine and Biology Society Annual Conference. 2005; 6: 5742–5745.
32. Peiris MTR, Davidson PR, Bones PJ, Jones RD. Detection of lapses in responsiveness from the EEG. Journal of neural engineering. 2011; 8 (1): 016003.
33. Davidson PR, Jones RD, Peiris MT. Detecting behavioral microsleeps using EEG and LSTM recurrent neural networks. Conference proceedings : Annual International Conference of the IEEE Engineering in Medicine and Biology Society IEEE Engineering in Medicine and Biology Society Annual Conference. 2005; 6: 5754–5757.
34. Golz M, Sommer D, Krajewski J. Prediction of immediately occurring microsleep events from brain electric signals. Current Directions in Biomedical Engineering. 2016; 2 (1).
35. Khatami R, Luca G, Baumann CR, et al. The European Narcolepsy Network (EU-NN) database. J Sleep Res. 2016; 25 (3): 356-364.
36. Scheer D, Schwartz SW, Parr M, Zgibor J, Sanchez-Anguiano A, Rajaram L. Prevalence and incidence of narcolepsy in a US health care claims database, 2008-2010. Sleep. 2019; 42 (7).
37. Finelli LA, Baumann H, Borbely AA, Achermann P. Dual electroencephalogram markers of human sleep homeostasis: correlation between theta activity in waking and slow-wave activity in sleep. Neuroscience. 2000; 101 (3): 523-529.
38. Klem GH, Lüders HO, Jasper H, Elger C. The ten-twenty electrode system of the International Federation. Electroencephalogr Clin Neurophysiol. 1999; 52 (3): 3-6.
39. Purcell SM, Manoach DS, Demanuele C, et al. Characterizing sleep spindles in 11,630 individuals from the National Sleep Research Resource. Nat Commun. 2017; 8: 15930.
40. Hertig-Godeschalk A, Skorucak J, Malafeev A, Achermann P, Mathis J, Schreier DR. Microsleep episodes in the borderland between wakefulness and sleep. submitted. 2018.
41. Burg JP. A new analysis technique for time series data. paper presented at Advanced Study Institute on signal Processing, NATO Enschede, Netherlands, 1968. 1968.
42. Olbrich E, Rusterholz T, LeBourgeois MK, Achermann P. Developmental Changes in Sleep Oscillations during Early Childhood. Neural Plast. 2017; 2017: 6160959.
43. Olbrich E, Achermann P. Oscillatory events in the human sleep EEG - detection and properties. Neurocomputing. 2004; 58: 129-135.

44. Olbrich E, Achermann P. Analysis of oscillatory patterns in the human sleep EEG using a novel detection algorithm. J Sleep Res. 2005; 14 (4): 337–346.
45. Olbrich E, Claussen JC, Achermann P. The multiple time scales of sleep dynamics as a challenge for modelling the sleeping brain. Philosophical Transactions of the Royal Society a-Mathematical Physical and Engineering Sciences. 2011; 369 (1952): 3884-3901.
46. Ahlstrom C, Nyström M, Holmqvist K, et al. Fit-for-duty test for estimation of drivers' sleepiness level: eye movements improve the sleep/wake predictor. Transportation research part C: emerging technologies. 2013; 26: 20-32.
47. Achermann P. Sleep. In: Akay M, ed. Wiley Encyclopedia of Biomedical Engineering. John WIley & Sons, Inc.; 2006.
48. Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Comput. 1997; 9 (8): 1735–1780.
49. Gers FA, Schmidhuber J, Cummins F. Learning to forget: Continual prediction with LSTM. Ninth International Conference on Artificial Neural Networks (Icann99), Vols 1 and 2. 1999; (470): 850-855.
50. Breiman L. Random Forests. Machine Learning. 2001; 45 (1): 5-32.
51. Vapnik V, Chervonenkis A. A note on one class of perceptrons. Automation and remote control. 1964; 25 (1): 103.
52. Graves A, Liwicki M, Fernandez S, Bertolami R, Bunke H, Schmidhuber R. A Novel Connectionist System for Unconstrained Handwriting Recognition. Ieee Transactions on Pattern Analysis and Machine Intelligence. 2009; 31 (5): 855-868.
53. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980. 2014.
54. Mayumi Oshiro T, Santoro Perez P, Baranauskas J. How Many Trees in a Random Forest? Machine learning and data mining in pattern recognition 2012; Berlin, Germany.
55. Galton F. Finger Prints London. Macmillan and Co. 1892.
56. Smeeton NC. Early history of the kappa statistic. Biometrics. 1985; 41: 795.
57. Powers DM. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. Journal of Machine Learning Technologies 2011; 2 (1): 37-36.
58. Altman DG, Bland JM. Diagnostic tests. 1: Sensitivity and specificity. BMJ. 1994; 308 (6943): 1552.
59. Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Phys Ther. 2005; 85 (3): 257-268.
60. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977: 159-174.
61. Danker-Hopfe H, Kunz D, Gruber G, et al. Interrater reliability between scorers from eight European sleep laboratories in subjects with different sleep disorders. J Sleep Res. 2004; 13 (1): 63-69.
62. Magalang UJ, Chen NH, Cistulli PA, et al. Agreement in the Scoring of Respiratory Events and Sleep Among International Sleep Centers. Sleep. 2013; 36 (4): 591-596.

63. Danker-Hopfe H, Anderer P, Zeitlhofer J, et al. Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard. J Sleep Res. 2009; 18 (1): 74-84.
64. Rosenberg RS, Van Hout S. The American Academy of Sleep Medicine inter-scorer reliability program: sleep stage scoring. J Clin Sleep Med. 2013; 9 (1): 81-87.
65. Malafeev A, Laptev D, Bauer S, et al. Automatic Human Sleep Stage Scoring Using Deep Neural Networks. Front Neurosci. 2018; 12: 781.
66. Stephansen JB, Olesen AN, Olsen M, et al. Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. Nature Communications. 2018; 9.
67. Olbrich S, Mulert C, Karch S, et al. EEG-vigilance and BOLD effect during simultaneous EEG/fMRI measurement. NeuroImage. 2009; 45 (2): 319–332.
68. Qian D, Wang B, Qing Y, et al. Bayesian Nonnegative CP Decomposition-based Feature Extraction Algorithm for Drowsiness Detection. IEEE transactions on neural systems and rehabilitation engineering : a publication of the IEEE Engineering in Medicine and Biology Society. 2016.
69. Shoorangiz R, Weddell SJ, Jones RD. Prediction of microsleeps from EEG: Preliminary results. Conference proceedings : Annual International Conference of the IEEE Engineering in Medicine and Biology Society IEEE Engineering in Medicine and Biology Society Annual Conference. 2016; 2016: 4650–4653.
70. Phan H, Andreotti F, Cooray N, Chen OY, De Vos M. SeqSleepNet: End-to-End Hierarchical Recurrent Neural Network for Sequence-to-Sequence Automatic Sleep Staging. IEEE Trans Neural Syst Rehabil Eng. 2019.
71. Mikkelsen KB, Ebajemito JK, Bonmati-Carrion MA, et al. Machine-learning-derived sleep-wake staging from around-the-ear electroencephalogram outperforms manual scoring and actigraphy. J Sleep Res. 2018: e12786.

Figure legends

Figure 1: *EEG (top) and corresponding spectrogram (bottom; Burg's algorithm; 1-s sliding window moved in steps of 200 ms) of derivation O2-M1. A 20-s epoch with a microsleep episode delineated by red lines is illustrated. Scaling of power density: -20 dB* *30 dB; 0 dB = 1 μV²/Hz.*

Figure 2: *Features used for the classification of microsleep episodes (MSEs). A 180-s segment is illustrated with the occurrence of MSEs indicated by the red shading. Features consist of power in the delta (0.8 – 4 Hz), theta (4 – 8 Hz), alpha (8 – 12 Hz), beta (12 – 26 Hz) frequency bands, the ratio theta/(alpha+beta) (T/AB), eye movements (delta activity of EOG divided by delta activity of O2-M1) and median EEG frequency (0.8 – 26 Hz range). Derivation O2-M1 was analyzed. Features were calculated for a 1-s sliding window moved in steps of 200 ms through the data. Power in the different bands was smoothed by a 1-s moving median filter.*

Figure 3: *Microsleep episodes (MSEs) of one patient scored by an expert (red) and detected by three classifiers (blue) are depicted. Long short-term memory (LSTM) neural network, random forest (RF), and support vector machine (SVM). As the training was performed only on MSEs and wakefulness (unilateral MSEs, MSE candidates and episodes of drowsiness were omitted), only MSEs and wakefulness are plotted in this figure, thus, the x-axis is compressed due to the omission of episodes. The entire recording of this patient is illustrated in supplementary Figure S3 (ID: tG6i).*

Figure 4: *Feature importance obtained in training (panels A and B) and in testing (panels C, D and E). A: Importance (arbitrary scale) of the different features used in the classification with the Random Forest (RF) approach. B: Performance (Cohen's kappa coefficient) of the LSTM network when one feature at a time was permuted in the training. C-E: Performance (Cohen's kappa coefficient) of the three classifiers when one feature at a time was permuted in the testing. Features were calculated for O1-M2 and O2-M1 leads and their combination was used for the training on 53 patients. For panels B-E one feature at a time was permuted in both O1-M2 and O2-M1 leads at the same time for testing in 23 patients. Higher values indicate a higher feature importance in panel A, while the opposite holds for panels B-E. See Figure 2 for the definition of the features.*

Tables

	Training	*Testing*
n	53	23
Male	38	12
Female	15	11
Age	46.4 ± 19.0	43.8 ± 15.2
Total # MSEs	912	351
Total duration MSE (min)	160.6	56.1
Total # MSEc	733*	231
Total duration MSEc (min)	37.1*	11.6
Total # ED	860*	392
Total duration ED (min)	102.4*	46.9
% sleep apnea patients	32.1	26.1
% EDS with unclear cause	30.2	43.5
% excessive tiredness	9.4	8.7
% narcolepsy	9.4	4.3
% idiopathic hypersomnia	5.7	4.3
% non-organic hypersomnia	1.9	0
% insomnia	1.9	0
% others	9.4	8.7

Table 1: *Demographic data, diagnosis, and total number and duration of MSE, MSEc and ED of patients contributing to the training and test data sets: total number of patients (n), number of males, females, mean age of patients and standard error of the mean, total number of MSEs, MSEc, and ED, total duration in minutes of MSEs, MSEc, and ED, and the percentage of patients with a suspected diagnosis of sleep apnea, EDS with unclear cause, excessive tiredness, narcolepsy, idiopathic hypersomnia, non-organic hypersomnia, insomnia and others. * not used for the training of the classifiers.*

A: *Only MSE and wakefulness considered*

	Sensitivity	*Specificity*	*Precision*	*Accuracy*	*Kappa*
LSTM	92.1	98.8	85.3	98.4	0.88
	87.7 ± 5.0	98.7 ± 0.5	85.7 ± 6.9	98.2 ± 0.5	0.83 ± 0.06
RF	90.7	98.5	81.8	98.0	0.85
	83.5 ± 5.8	98.2 ± 0.7	81.7 ± 8.2	97.7 ± 0.7	0.78 ± 0.08
SVM	88.0	98.1	77.2	97.4	0.81
	80.4 ± 5.7	97.8 ± 1.0	80.5 ± 9.1	97.1 ± 1.0	0.75 ± 0.08

B: *Unilateral MSE, MSEc, and ED considered as wakefulness*

	Sensitivity	*Specificity*	*Precision*	*Accuracy*	*Kappa*
LSTM	92.1	96.8	66.0	96.5	0.75
	87.8 ± 4.9	96.1 ± 0.2	63.9 ± 7.0	96.2 ± 0.2	0.68 ± 0.06
RF	89.4	97.0	66.8	96.5	0.75
	81.9 ± 6.6	96.2 ± 0.2	63.8 ± 7.7	96.2 ± 0.2	0.66 ± 0.07
SVM	87.0	96.7	63.7	96.0	0.71
	79.7 ± 6.2	95.9 ± 0.2	62.9 ± 8.4	95.7 ± 0.3	0.63 ± 0.08

C: *Unilateral MSE, MSEc, and ED considered as MSEs*

	Sensitivity	*Specificity*	*Precision*	*Accuracy*	*Kappa*
LSTM	55.7	98.8	88.4	92.8	0.65
	53.3 ± 8.1	98.7 ± 0.5	81.1 ± 7.8	92.3 ± 1.7	0.58 ± 0.06
RF	51.7	98.6	85.3	92.0	0.60
	50.7 ± 9.1	98.3 ± 0.7	77.8 ± 10.5	91.6 ± 2.0	0.55 ± 0.08
SVM	50.1	98.1	81.0	91.4	0.57
	54.0 ± 7.7	97.9 ± 1.0	84.1 ± 8.7	90.9 ± 2.1	0.52 ± 0.08

D: *Unilateral MSE and MSEc considered as MSEs, ED considered as wakefulness*

	Sensitivity	*Specificity*	*Precision*	*Accuracy*	*Kappa*
LSTM	81.2	97.6	75.0	96.3	0.76
	71.5 ± 8.4	97.1 ± 0.7	67.2 ± 8.1	96.1 ± 0.9	0.64 ± 0.08
RF	77.7	97.7	74.8	96.1	0.74
	72.5 ± 7.4	97.1 ± 0.9	73.7 ± 7.3	95.8 ± 1.0	0.66 ± 0.07
SVM	75.7	97.3	71.4	95.5	0.71
	70.7 ± 6.9	96.8 ± 1.1	72.4 ± 8.0	95.3 ± 1.2	0.64 ± 0.08

Table 2: *Performance of the classifiers in percentages for all measures except for kappa. Performance measures: sensitivity, specificity, precision, accuracy and Cohen's kappa coefficients. Three classification algorithms were evaluated in 23 patients (test set): long short-term memory (LSTM) neural network, random forest (RF), and support vector machine (SVM). Overall performance measures across all patients (gray shading; data of all patients were pooled) and mean across patients and standard error of the mean (white). Recordings not having MSEs (≤1 MSE, n=13) were excluded for calculation of mean performance for sensitivity, precision and Cohen's kappa since these measures take into account positives (i.e. MSEs). The performance was calculated based on the 200-ms resolution.* ***A:*** *algorithm performance taking into account only MSEs and wakefulness;* ***B:*** *unilateral MSEs, MSEc, and ED were assigned to the category wakefulness;* ***C:*** *in addition to MSEs, unilateral MSEs, MSEc, and ED were considered as MSEs,* ***D:*** *unilateral MSEs and MSEc were assigned to the category MSEs, while ED were assigned to wakefulness.*

	Sensitivity	*Specificity*	*Precision*	*Accuracy*	*Kappa*
	74.1	98.1	86.6	94.7	0.77
	62.6 ± 12.7	97.7 ± 1.6	91.1 ± 2.8	94.7 ± 1.5	0.67 ± 0.10

Table 3: *Inter-scorer performance (5 patients; 2 independent scorers). The performance was calculated with the 200-ms resolution. Unilateral MSE, MSEc and ED were assigned to the category wakefulness for calculating the inter-scorer performance.*

Figure 1

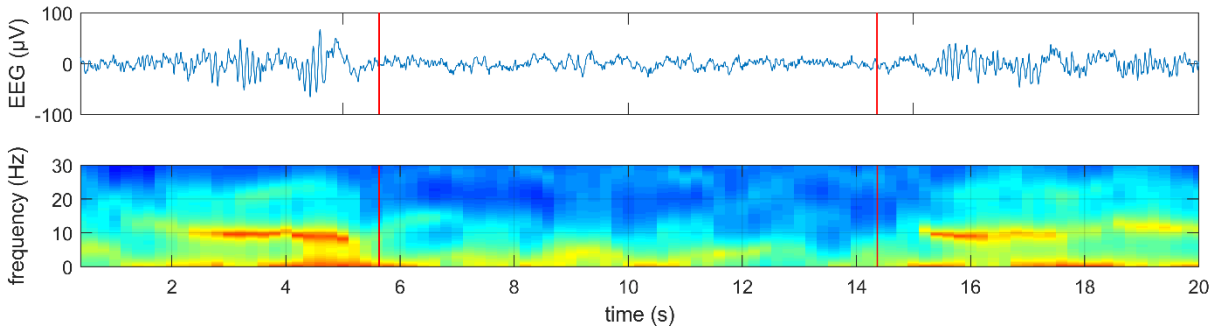

Figure 2

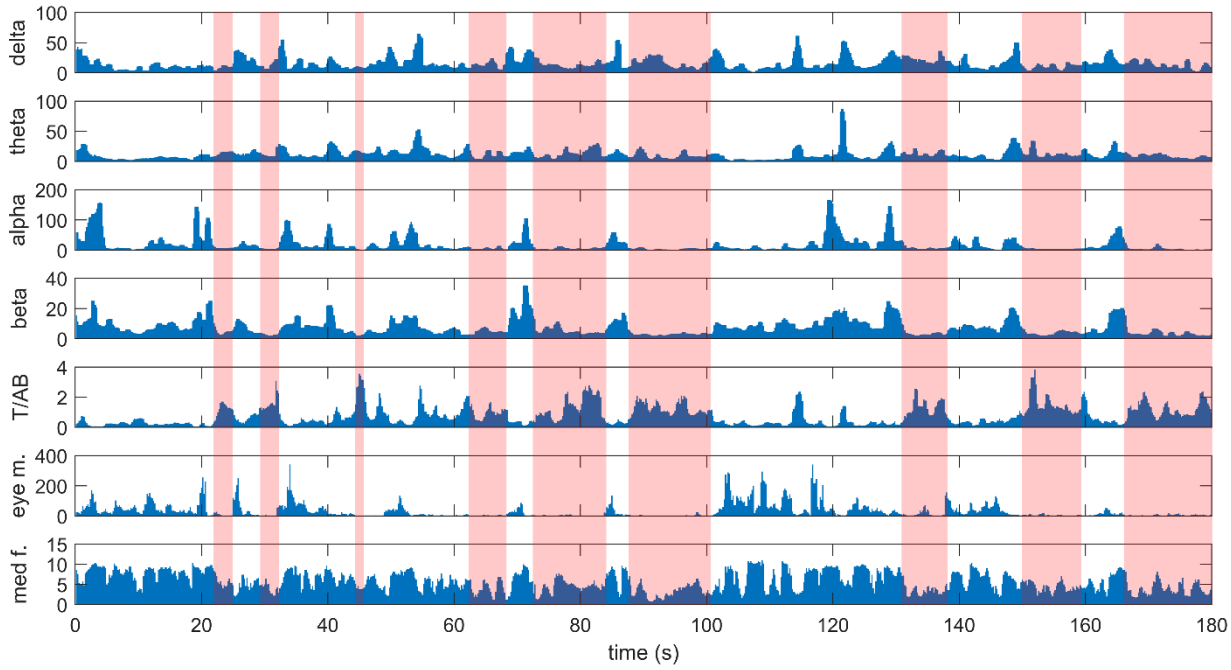

Figure 3

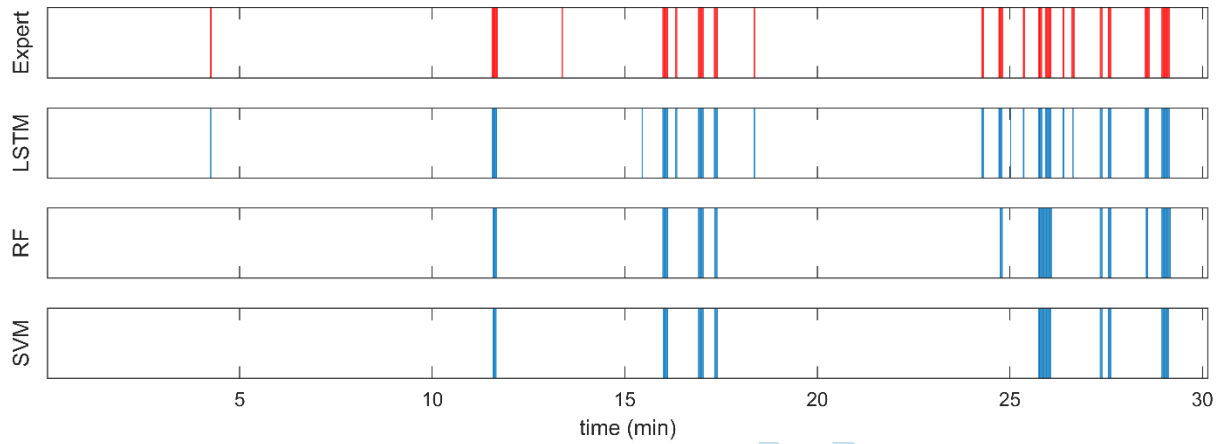

Figure 4