Open Access

Martin Golz*, David Sommer and Jarek Krajewski

Prediction of immediately occurring microsleep events from brain electric signals

DOI 10.1515/cdbme-2016-0035

Abstract: This contribution addresses the question if imminent changes of the cortical state are predictable. The analysis is based on 1484 examples of microsleep (MS) and 1940 counterexamples of sustained attention (SA), both observed during overnight driving in the simulator. EEG segments (8 s in length) immediately before each respective event were included. Features were extracted by (i) modified periodogram and (ii) Choi-Williams distribution. Machine learning algorithms, namely the optimized learning vector quantization (OLVQ) and the support-vector machine with Gaussian kernel function (SVM), were trained in order to map signal features to the event type (MS or SA). Cross validation analysis yielded test set classification accuracies of 87.5 $\pm$ 0.1 % and 82.7 $\pm$ 0.1 % for feature set (i) and (ii), respectively. In general, SVM outperformed OLVQ. In conclusion, EEG contains enough information to predict immediately upcoming microsleep events.

Keywords: driving simulation; EEG; microsleep; neural networks; sleepiness; support-vector machines.

1 Introduction

The detection of short-term brain states based on biosignals is still challenging. One important example of such brain states are attention losses during driving or during other tasks that require sustained attention. Main causes of short and unintended attention losses are sleepiness and monotony. Thus, they are often called microsleep (MS) events. It has been shown that the detection of MS from electroencephalographic and electrooculographic recordings is possible if computational intelligence algorithms, like artificial neural networks or margin optimizing algorithms within the recursive kernel Hilbert space, i.e. support-vector machines (SVM), are applied [1, 2]. To the best of our knowledge, the short term prediction of MS events has not been investigated up to now.

In a recent paper, sleepiness estimations were presented based solely on EEG utilizing time-frequency analysis, especially the Choi-Williams distribution [3]. Here, the same methods were applied to describe sleepiness, but within a short-time arrangement in order to predict MS.

2 Material and methods

2.1 Experimental design

Data were recorded during a driving simulation study in the year of 2007 in cooperation with Caterpillar Machine Research Inc. Sixteen young adults (12 male, 4 female, mean age 24.4 $\pm$ 3.1 years) were randomly selected out of 133 interested volunteers. Two experimental nights per subject were conducted. During three days before experiments, the scheduled sleep-wake restrictions (wake-up time: 6:00 – 8:00, time to bed: 22:00 – 1:00) were monitored by wrist actimetry. Furthermore, it was verified that subjects refrained from naps during the day before driving. This way, a relatively large time since sleep was ensured which is an important factor of sleepiness. Experiments started at 22:30 and ended at 7:30 and included 8 driving sessions of 40 min duration each. EEG (Fp1, Fp2, A1, C3, Cz, C4, A2, O1, O2; common average reference) and EOG (vertical, horizontal) were recorded (device: Sigma PLpro, Sigma Medizintechnik GmbH, Gelenau, Germany). Further details were published earlier [4].

Behavioural MS events had been evaluated by an expert supervisor. Evaluations were based on visual inspections of video material, of lane deviation time series and of EOG. Vague or very short-lasting signs of MS (< 0.3 s) were excluded. The duration of behavioural signs of MS ranged between 0.3 and 6 s. On the other hand, periods of time where the driver is indeed drowsy, but still able to keep

***Corresponding author: Martin Golz,** University of Applied Sciences, 98574 Schmalkalden, Germany, E-mail: m.golz@hs-sm.de
David Sommer: University of Applied Sciences, 98574 Schmalkalden, Germany, E-mail: sommer.david@gmail.com
Jarek Krajewski: University of Applied Sciences Cologne, Germany, E-mail: krajewski@rfh-koeln.de

the car in lane, were used as counterexamples and were labelled as sustained attention (SA).

2.2 Biosignal processing

Two different cases of segmentation were considered:

1. Prediction: the whole EEG segment is located before an MS or SA event,
2. Detection: the EEG segment is centred around an MS or SA event (see Figure 1).

A practical application could process such EEG data in real time and decide immediately and continuously whether there is an MS event or not. The decision would be too late in the detection case and just in time in the prediction case. The EEG segment length was fixed to 8 s.

Power spectral densities (PSD): Under the assumption that the EEG is quasi-stationary within the segment, PSD values were estimated based on the periodogram method modified by data tapering (Hann taper). PSD were logarithmically scaled (in decibel units) and band-averaged in small frequency bands between 0.1 and 23.1 Hz. The width of these bands was empirically optimized, resulting in 1 Hz. As alternative estimation methods, the weighted overlapped segment analysis (Welch's method) and the multi-taper method (Thomson's method) were applied, but yielded slightly less accuracy in subsequent machine learning.

Choi-Williams distribution (CWD): Assuming that the EEG is non-stationary within the segment, then analyses by methods of the Cohen class might be considered. Following the suggestions of [3], the CWD was utilized. CWD satisfies a number of mathematical properties. For signals with quasi-constant instantaneous frequency it achieves potentially better time-frequency concentration than other methods, e.g. the smoothed pseudo Wigner-Ville distribution, which are more accurate for chirp-like signals. CWD depends on the signal's time-frequency structure and performs potentially poor if signal components overlap in time and frequency. The cross-CWD between two signals $x(t)$ and $y(t)$ is expressed as follows [5]:

$$T_{xy}(t,f) = 2\iint_{-\infty}^{\infty} \frac{\sqrt{\sigma}}{4\sqrt{\pi}\,|\tau|} e^{-\nu^2\sigma/(16\tau^2)} \cdots x\left(t+\nu+\frac{\tau}{2}\right) y^{\star}\left(t+\nu-\frac{\tau}{2}\right) e^{-j2\pi f\tau}\, d\nu\, d\tau \quad (1)$$

If $x(t) = y(t)$, then the auto-CWD of $x(t)$ is estimated. The kernel width σ was fixed to 0.005 [3]. For each of the four EEG bands ($\delta = 0.1 - 4$ Hz, $\theta = 4 - 8$ Hz, $\alpha = 8 - 12$ Hz, and $\beta = 12 - 35$ Hz), the instantaneous power function (IP) was estimated:

$$S_{xy,i}(t) = \frac{\int_{f_{i1}}^{f_{i2}} |T_{xy}|\, df}{\int_{f_{TB1}}^{f_{TB2}} |T_{xy}|\, df} \quad (2)$$

$S_{xy,i}$ denotes the IP of the i-th band, f_{i1} and f_{i2} the lower and upper band limit of the i-th band, and f_{TB1} and f_{TB2} the lower and upper band limit of the total band. In accordance to [3], these limits were set to f_{TB1} = 0.1 Hz and f_{TB2} = 45 Hz. The instantaneous frequency (IF) was estimated as the first moment of the CWD with respect to frequency:

$$f_{IF}(t) = \frac{\int_{TB1}^{TB2} f T_{xy}(t,f)\, df}{\int_{f_{TB1}}^{f_{TB2}} T_{xy}(t,f)\, df} \quad (3)$$

Finally, the complexity of the CWD within each band was estimated by means of the Shannon entropy:

$$H_i(t) = -\int_{f_{i1}}^{f_{i2}} T_{xy}(t,f)\, \log_2\left(T_{xy}(t,f)\right) df \quad (4)$$

High values of the instantaneous spectral entropies (ISE) indicate broad-band spectral characteristics (uniform, flat spectrum), whereas low ISE values imply line spectrum characteristics, with power concentrated into a few frequency bins.

From each of these functions (eqs. 2 – 4), the following parameters of descriptive statistics were calculated: mean, median, minimum, maximum [3]. This led to 4 parameters of the IP functions (eq. 2) in 4 frequency bands, to 4 parameters of the IF function (eq. 3), and to 4 parameters of the ISE functions (eq. 4) in 4 frequency bands.

In accord with [3], feature extraction was performed on

1. Auto-CWD of the following 7 channels: Fp1, Fp2, C3, Cz, C4, O1, O2;
2. Cross-CWD of the following 6 ipsilateral pairs: Fp1-C3, Fp2-C4, Fp1-O1, Fp2-O2, C3-O1, C4-O2;

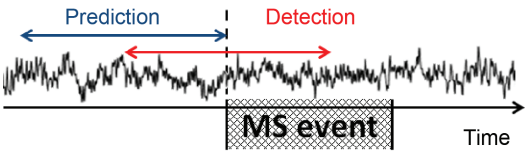

Figure 1: Two cases of EEG segmentation around each MS or SA event.

3. Cross-CWD of the following 5 contralateral pairs: Fp1-Fp2, C3-C4, O1-O2, C3-O2, C4-O1.

This resulted in 36 features in each of 18 channels and channel pairs, i.e. 648 CWD features per EEG segment. In case of PSD, 23 features were extracted per channel (see above). For these preliminary investigations, cross spectral densities were not estimated. Therefore, only 7 channels, but no channel pairs, were included in PSD estimation. This generated 161 PSD features per EEG segment.

2.3 Classification analysis

The goal of this step is to find a mapping of high-dimensional feature vectors $\boldsymbol{g} \in \mathbb{R}^d$ to class labels $c \in \mathbb{Z}$. Here, the feature space had a dimensionality of $d = 648$ and $d = 161$ for CWD and PSD, respectively. Classification was restricted to a two-class problem: $c = +1$ for MS events and $c = -1$ for SA behaviour of the driver.

Optimized learning vector quantization (OLVQ1): [6] is an artificial neural network which performs stochastic learning during which weight vectors $\boldsymbol{w_k} \in \mathbb{R}^d$, $k = 1, \ldots, n_N$ are adapted to regions of high probability densities. It aims at approximating the probability density function in feature space. The number of neurons n_N is the most important design parameter of OLVQ1. With large n_N the approximation is finer than with low n_N. Therefore, with increasing n_N, OLVQ1 tends to be more adaptive and results in higher classification performance. Amongst several LVQ algorithms, OLVQ1 has an adaptive step size parameter η_k for each individual neuron ($k = 1, \ldots, n_N$). If a neuron converges badly during training iterations, then η_k is increased, otherwise it is decreased. For many practical cases it has been shown by several authors that OLVQ1 outperforms other LVQ algorithms at the cost of a higher demand on main memory. In general, OLVQ1 converges rapidly in most applications.

Support-vector machines (SVM): Given a finite training set $\left\{(\boldsymbol{g}_i, c_i) \,|\boldsymbol{g}_i \in \mathbb{R}^d, c_i \in \mathbb{R}, i = 1, \ldots, N\right\}$ with fixed pairs of feature vectors $\boldsymbol{g}_i$, class labels c_i, and sample size N, one wants to find amongst all possible linear separation functions $\boldsymbol{w}^{\mathrm{T}}\,\boldsymbol{g} + b = 0$, that which maximizes the margin, i.e. the distance between the linear separation function and the nearest feature vectors of each class. This optimization problem is solved at the saddle point of the Lagrange functional [7]:

$$L(\boldsymbol{w}, b, \alpha) = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{N} \alpha_i \left(c_i \left(\boldsymbol{w}\,\boldsymbol{g}_i + b\right) - 1\right) \qquad (5)$$

using Lagrange multipliers α_i. Both the vector $\boldsymbol{w}$ and scalar b are to be optimized. The solution of (5) is given by

$$\boldsymbol{w}_S = \sum_{i=1}^{N} \alpha_i c_i \boldsymbol{g}_i \;;\; b_S = -\frac{1}{2}\boldsymbol{w}_S \left(\boldsymbol{g}_+ + \boldsymbol{g}_-\right) \qquad (6)$$

where $\boldsymbol{g}_+$ and $\boldsymbol{g}_-$ are support vectors with $\alpha_+ > 0$, $c_+ = +1$ and $\alpha_- > 0$, $c_- = -1$, respectively. If the problem is not solvable error-free, then a penalty term $\sum_{i=1}^{N_{sl}} \xi_i$ with N_{Sl} slack variables $\xi_i \geq 0$ has to be used which controls the amount of allowed classification errors [7]. This leads to a restriction of the Lagrange multipliers to the interval $0 \leq \alpha_i \leq C \;\forall i = 1, \ldots, N_{sl}$. The regularization parameter C has to be optimized empirically by minimizing the mean classification error of the given training set.

In order to extend the separation functions to nonlinear ones the SVM is extended by kernel functions $k(\boldsymbol{g}_i, \boldsymbol{g})$ with $\boldsymbol{g}_i$ as a feature vector of the given training set and $\boldsymbol{g}$ as a free test vector of the feature space:

$$\sum_{i=1}^{N} \alpha_i c_i k(\boldsymbol{g}_i, \boldsymbol{g}) + b = 0 \qquad (7)$$

The Gaussian radial basis function $k(\boldsymbol{g}_i, \boldsymbol{g}) = exp\left(-y\|\boldsymbol{g}_i - \boldsymbol{g}\|^2\right)$ had been used, where y is a free parameter which had been optimized empirically.

3 Results

As mentioned above, all parameters of OLVQ1 and SVM were optimized empirically in order to get highest mean classification accuracy during training. Most importantly for OLVQ1 is the number of neurons n_N which was varied between 2 and 400 (Figure 2). At least 50 neurons should be used, higher values lead to minor increases in mean test set classification accuracies.

Most importantly for SVM are the kernel width y and the regularization parameter C. It turned out, that there is a relatively small interval for γ ($10^{-5} \leq y \leq 10^{-3}$) where classification performed well for test data, i.e. data which were never presented during training (Figure 3). If both have been optimized then training set classification accuracies of 100% are achievable. Despite this perfect training, mean test set accuracies were never higher than 87.5%. For these calculations a huge amount of computational power was necessary, because of very low

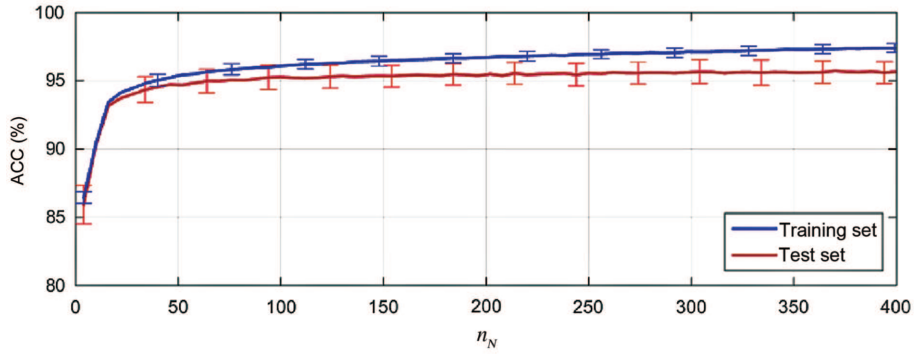

Figure 2: Mean and standard deviation of classification accuracy across the number of neurons n_N of the OLVQ1 classifier applied to the PSD feature set of the detection case.

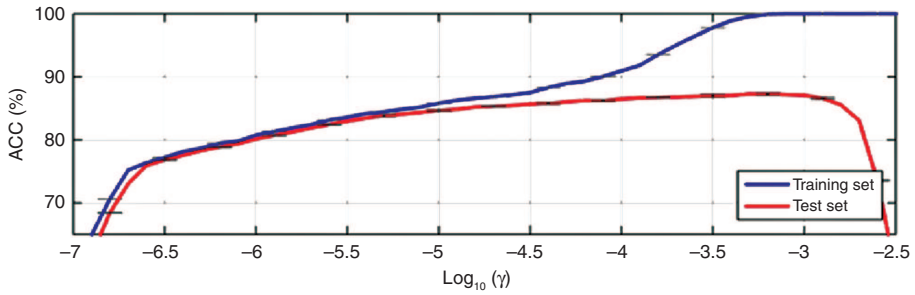

Figure 3: Mean and standard deviation of classification accuracy across the kernel width parameter γ of the SVM classifier applied to the PSD feature set of the prediction case.

Table 1: Classification accuracy (mean and standard deviation) for training data sets for the prediction and detection case based on two classifiers (OLVQ1, SVM) and two feature sets (CWD, PSD).

Classifier Feature set	OLVQ1 CWD	OLVQ1 PSD	SVM CWD	SVM PSD
Prediction	86.6 ± 0.5%	89.6 ± 0.5%	97.3 ± 0.0%	98.9 ± 0.0%
Detection	94.1 ± 0.5%	97.6 ± 0.3%	98.3 ± 0.0%	99.8 ± 0.0%

Table 2: Classification accuracy for the **test data sets**.

Classifier Feature set	OLVQ1 CWD	OLVQ1 PSD	SVM CWD	SVM PSD
Prediction	81.6 ± 1.5%	82.9 ± 1.5%	82.7 ± 0.1%	87.5 ± 0.1%
Detection	91.1 ± 1.3%	95.5 ± 0.7%	91.2 ± 0.1%	97.3 ± 0.1%

convergence of the SVM training. In a grid computing framework, more than 40 high-performance workstations performed distributed learning and distributed parameter optimization.

Results of training for empirically optimized parameter set tings are summarized in Table 1. OLVQ1 was always out performed by SVM and in all cases PSD features led to higher training accuracy than CWD features. With mean accuracy between 97.3 % and 99.8 % it was possible to gain nearly perfect training results with SVM.

High training success is essential to gain the essence of machine learning: generalizability, i.e. future data should be classified accurately. This is verified by test data which were never presented during training. Results offer slight decreases in mean accuracies (Table 2) compared to mean accuracies of training data (Table 1). Again, OLVQ1 has been always outperformed by SVM, and PSD features always led to higher accuracy than CWD features. Comparing both segmentation cases, it turned out that for segments of the detection case MS events are much more accurately discriminable from SA behaviour than for segments of the prediction case.

4 Discussion and conclusion

The Choi-Williams distribution is a time-frequency analysis method which has been demonstrated to perform

well for signals composed of several oscillatory processes with constant or varying instantaneous frequency in the presence of low noise. EEG is known as a composition of a huge amount of stochastic processes which mostly result in spectra with broad-band characteristics. Nevertheless, it is an open question how useful CWD performs for such processes. Here and also in [3], visual inspection of CWD results was replaced by multivariate analysis. Mean accuracies of about 75% were obtained from 20 subjects with excessive daytime sleepiness compared to 20 subjects without daytime sleepiness [3]. Here, a mean accuracy of 82.7% for CWD and 87.5% for PSD was obtained on records of 14 subjects. In contrast to [3] a short-term prediction of immediately upcoming events (MS or SA) during sleepiness was assessed. If it is allowed to observe the already existing event (detection case) then a much higher mean accuracy (97.3%) has been obtained.

Despite the large amount of noise and the large variance of the periodogram estimation, it seems that machine learning methods get enough information from multichannel EEG to predict correctly immediately upcoming microsleep events. It is an open question how robust the results are if the sample size is increased by EEG of subjects not included in the training process of the machine learning methodology.

Acknowledgment: The authors would like to thank Mr. Adolf Schenka for technical support during the lab study and for data management.

Author's Statement

Research funding: The author state no funding involved. Conflict of interest: Authors state no conflict of interest. Informed consent: Informed consent has been obtained from all individuals included in this study. Ethical approval: The research related to human use complies with all the relevant national regulations, institutional policies and was performed in accordance with the tenets of the Helsinki Declaration, and has been approved by the authors' institutional review board or equivalent committee.

References

[1] Davidson PR, Jones R, Peiris M. EEG-based lapse detection with high temporal resolution. IEEE Trans Biomed Eng. 2007;54:832–9.

[2] Golz M, Sommer D, Chen M, Trutschel U, Mandic D. Feature fusion for the detection of microsleep events. J VLSI Signal Proc. 2007;49:329–42.

[3] Melia U, Guaita M, Vallverdú M, Clariá F, Montserrat JM, Vilaseca I, et al. Characterization of daytime sleepiness by time-frequency measures of EEG signals. J Med Biol Eng. 2015;35:406–17.

[4] Golz M, Sommer D, Trutschel U, Sirois B, Edwards D. Evaluation of fatigue monitoring technologies. Somnologie. 2010;14:187–99.

[5] Choi HI, Williams WJ. Improved time-frequency representation of multicomponent signals using exponential kernels. IEEE T Acoust Speech. 1989;37:862–71.

[6] Kohonen T. Self-organizing maps. Berlin: Springer; 2003.

[7] Vapnik VN. Statistical learning theory. N.Y.: Wiley; 1998.