Martin Golz*, Adolf Schenka, Florian Haselbeck and Martin P. Pauli

Inter-individual variability of EEG features during microsleep events

<https://doi.org/10.1515/cdbme-2019-0004>

Abstract: This paper examines the question of how strongly the spectral properties of the EEG during microsleep differ between individuals. For this purpose, 3859 microsleep examples were compared with 4044 counterexamples in which drivers were very drowsy but were able to perform the driving task. Two types of signal features were compared: logarithmic power spectral densities and entropy measures of wavelets coefficient series. Discriminant analyses were performed with the following machine learning methods: support-vector machines, gradient boosting, learning vector quantization. To the best of our knowledge, this is the first time that results of the leave-one-subject-out cross-validation (LOSO CV) for the detection of microsleep are presented. Error rates lower than 5.0 % resulted in 17 subjects and lower than 13 % in another 11 subjects. In 3 individuals, EEG features could not be explained by the pool of EEG features of all other individuals; for them, detection errors were 15.1 %, 17.1 %, and 27.0 %. In comparison, cross validation by means of repeated random subsampling, in which individuality is not considered, yielded mean error rates of 5.0 ± 0.5 %.

A subsequent inspection of raw EEG data showed that in two individuals a bad signal quality due to poor electrode attachment could be the cause and in one individual a very unusual behavior, a high and long-lasting eyelid activity which interfered the recorded EEG in all channels.

Keywords: Microsleep, EEG, periodogram, support-vector machines, SVM, learning vector quantization, LVQ, gradient boosting, leave one out, LOSO cross-validation.

<https://doi.org/10.1515/cdbme-2019-0004>

1 Introduction

The recognition of changes in brain state is a challenging task. This is especially the case for short-term requirements, e.g.

when microsleep events (MSE) of car drivers should be detected from the EEG. Machine learning methods have been successfully used for this purpose [1,2]. Because these methods are based solely on the given data set and on assumptions about the underlying data generating, stochastic processes, great care must be taken when selecting data [3]. Therefore, one research focus should be in validation. On the one hand, training sample size should be as large as possible in order to enable confident and accurate learning, and on the other hand enough test examples should be available, which must be drawn identically and statistically independent from the same unknown distribution as the training examples [4].

To the best of our knowledge, this is the first time that results of a leave-one-subject-out cross-validation (LOSO CV) [4] for MSE detection are presented. LOSO CV involves training on data of all subjects except one and validation on data of the subject held out. This procedure is performed for each individual such that data of every individual was $n_I - 1$ times used for training and at one time used for testing, whereby n_I is the number of individuals. By holding out data of an individual, LOSO CV simulates the case that data of the individual are not currently available for machine learning training, but in the future for testing the learned model. It asks how much the results are influenced if data of an individual are available or not and therefore, it asks how general the results of machine learning can be interpreted.

2 Material

LOSO CV analysis is based on data of two studies performed in our driving simulation lab in the years 2007 [5] and 2016. The first was completed by 16 (5 ♀, 11 ♂, age: 24.4 ± 3.1) and the second by 15 young adults (7 ♀, 8 ♂, age: 24.4 ± 2.8). The procedure was almost the same in both studies: seven driving sessions with a duration of 40 minutes were started hourly between 1 and 8 am. All subjects wore an actometer for three days before the study, so that it could be tested, among other things, that the end of sleep was at least 8 am and that subjects initiated sleep latest at 1 am in the two nights before the study. The study design of both studies ensured that the following

*Corresponding author: **Martin Golz:** University of Applied Sciences Schmalkalden, Blechhammer 4, Schmalkalden, Germany, e-mail: m.golz@hs-sm.de

Adolf Schenka, Florian Haselbeck, Martin Patrick Pauli: University of Applied Sciences, Schmalkalden, Germany

four factors were effective for achieving high drowsiness: (1) *Time since sleep* was at least 16 hours, (2) *Time on task* was relatively long (280 min), (3) *Time of day* was near the circadian trough, and (4) *High monotony* occurred due to the driving task and the absence of communication with others.

There were differences between the both studies, particularly in the technical lab equipment, including the EEG devices. The SigmaPL-Pro® (Neurowerk GmbH, Gelenau, Germany) was used in 2007 and the SomnoScreen® (Somnomedics GmbH, Kist, Germany) in 2016. Electrodes were attached to eight positions (Fp1, Fp2, C3, C4, O1, O2, A1, A2; reference electrode: Cz; common average reference).

Based on video recordings of the eye region, head and shoulders as well as the driving scene, the starting times and lengths of MSE were determined. MSE have always been determined based on observable behavioral characteristics, in particular prolonged eyelid closure and slow eye movement. This visual assessment was performed by a trained person with long experience in this field [3]. A total of 7903 events consisting of 3859 MSE and of 4044 counterexamples was drawn from recordings.

3 Methods

The machine learning methods achieved lowest error rates if the EEG was segmented such that the beginning and end of the EEG segments were 1 s before and 3 s after the start time of MSE occurrence. This setting was fixed uniformly for all channels and all subjects. Details of these empirical optimizations are addressed in section 4. Logarithmic power spectral densities (LogPSD) were estimated using the modified periodogram with Hann tapers. Alternatively, the following direct and indirect PSD estimation methods were tested: Welch, multi-taper, Yule-Walker, Burg. For them, however, the final error rates of machine learning were slightly higher. Additionally, feature extraction was extended to the discrete wavelet transform (DWT) and to wavelet packet transform. Signal decomposition was performed up to level 8. The total power and the following entropies were estimated for each detail and for the approximation coefficient series: Shannon, threshold, SURE, l_p norm, and mean logarithmic instantaneous power [6]. The DWT using Daubechies mother wavelet of order 2 (db2) at decomposition level 5 and with the total power as well as all 5 types of entropies provided feature sets that proved most successful for machine learning. Thus, from each EEG segment a 216-dimensional feature vector consisting of 6 measures (total power and 5 entropies) of 6 wavelet coefficient series (1 approximation, 5 details) in each of 6 EEG channels

(A1 and A2 channels were not processed) was extracted. Comparisons of the machine learning performance for these DWT feature sets and for the LogPSD feature sets will be presented in section 4.

Empirical optimizations of LogPSD estimates led to the following setting: spectral bands having a width of 1.0 Hz across the interval from 0.2 to 40.2 Hz. This way, the number of LogPSD values was 40 per channel and 240 in total. Processing a smaller number of channels led to increased errors.

For supervised training of machine learning methods, a sample containing pairs of feature vectors $\mathbf{x}_n$ and target variables t_n must be provided:

$$\mathcal{S} = \{(\mathbf{x}_n, t_n) | \mathbf{x}_n \in \mathbb{R}^d, t_n \in \{0, 1\}, n = 1, 2, \dots, N\} \quad (1)$$

with $N = 7903$; $d = 240$ for LogPSD, $d = 216$ for DWT. As binary target variable the class label $t_n = 1$ was used for MSE examples and $t_n = 0$ for counterexamples.

The following three different methods of machine learning have been applied in order to compare their performance:

- Learning vector quantization (LVQ)
- Gradient boosting (GB)
- Support-vector machines (SVM)

LVQ is a neural net consisting of one layer of neurons performing competitive learning. I.e., the neuron whose weight vector $\mathbf{w}_k$ is closest to the current input vector $\mathbf{x}_n$ in terms of a pre-selected vector norm wins the competition among all neurons and is adapted by the following learning rule:

$$\Delta \mathbf{w}_c = \pm \eta (\mathbf{x}_n - \mathbf{w}_c) \quad (2)$$

with step size $\eta \in (0, 1)$, $\mathbf{w}_c = \operatorname{argmin} \|\mathbf{x}_n - \mathbf{w}_k\|$, $\mathbf{x}_n \in \mathcal{S}$, $\mathbf{w}_k \in \{(\mathbf{w}_k, \tau_k) | \mathbf{w}_k \in \mathbb{R}^d, \tau_k \in \{0, 1\}, k = 1, \dots, K, K \ll N\}$.

The gradient boosting (GB) algorithm creates an ensemble of decision trees. At each iteration, for each class a new tree is created such that they correct the errors of the former tree. This is fulfilled when the squared error loss function has a negative gradient. Trees added to the ensemble are no more modified. At the end, for a given data example all trees are retrieved and a majority decision is made: The class that was calculated by more trees is selected. Since GB tends to overfit relatively quickly, the decision trees should be limited according to the number of branches per level and the depth of the tree. Here LightGBM was used, a numerically efficient GB variant developed by Microsoft Inc. [7], which uses existing graphics processing unit in addition to the central processing unit. However, our data sets were too small to benefit from using the graphics processing unit. LightGBM tends to have deeper rather than wider decision trees [7].

SVM aims at finding a mapping $\mathbf{x} \mapsto t$ for any vector $\mathbf{x} \in \mathbb{R}^d$ based on all samples $(\mathbf{x}_n, t_n) \in S$. The largest possible margin of a linear separation function $\mathbf{w}^T \mathbf{x} + b = 0$ between the two class domains is sought. I.e. the parameters $\mathbf{w} \in \mathbb{R}^d, b \in \mathbb{R}$ must be optimized provided that the distance between the separation function and the nearest feature vectors, the support vectors, is maximum. It has been proved that this problem has a unique solution. The corresponding Lagrangian:

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{n=1}^N \alpha_n (t_n (\mathbf{w}^T \mathbf{x}_n + b) - 1) \quad (3)$$

must be in a saddle point because $L(\mathbf{w}, b, \alpha)$ must be minimized with respect to $\mathbf{w}, b$ and maximized with respect to α . The solution can be formulated explicitly:

$$\mathbf{w}_S = \sum_{n=1}^N \alpha_n t_n \mathbf{x}_n, \quad b_S = t_n - \mathbf{w}_S^T \mathbf{x}_n \quad (4)$$

Only the set of support-vectors $\{\mathbf{x}_n | \alpha_n > 0\}$ contribute to equ. (4). If the training set is not separable, an error term $C = \sum_{i=1}^{N_{sl}} \xi_i$ with N_{sl} different slack variables $\xi_i \geq 0$ can be introduced to forgive classification errors (soft margin principle). This leads to a restriction of the multipliers to the interval $0 \leq \alpha_n \leq C, \forall n = 1, \dots, N$. The regularization parameter C must be optimized empirically by minimizing mean training errors. Additionally, the solution should be searched within the high-dimensional reproducing kernel Hilbert space by applying an admissible kernel function $k(\mathbf{x}_n, \mathbf{x})$ in order to take advantage of the blessings of high dimensionality and to get non-linear separation functions in the input space. This way, the separation function changes from $\mathbf{w}^T \mathbf{x} + b = 0$ to $k(\mathbf{x}_n, \mathbf{x}) + b = 0$ and the equation (3) changes to $L(\mathbf{w}, b, \alpha) = 1/2 \|\mathbf{w}\|^2 - \sum_{n=1}^N \alpha_n (t_n (k(\mathbf{x}_n, \mathbf{x}) + b) - 1)$. Gaussian kernel functions $k(\mathbf{x}_n, \mathbf{x}) = \exp(-\gamma \|\mathbf{x}_n - \mathbf{x}\|^2)$ were used, which have only one free parameter γ to be optimized empirically.

Figure 1: Mean and standard deviation of classification errors on the test set versus the time offset between segment center and MSE start time. Results were estimated using the LVQ algorithm on LogPSD and on DWT feature sets.

Table 1: Mean and standard deviation of *MSE detection errors* estimated by repeated random subsampling using three different learning methods. The number of EEG segments N and the number of LogPSD features N_F were equal for all methods.

Method	N	N_F	E_{TRAIN} [%]	E_{TEST} [%]
LVQ	7,903	240	2.42 ± 0.14	6.45 ± 0.58
GB	7,903	240	0.00 ± 0.00	5.10 ± 0.40
SVM	7,903	240	0.19 ± 0.03	5.03 ± 0.53

4 Results

First, results of the repeated random subsampling are shown. It turned out that the length of EEG segments can be varied between 4 and 10 s without any degradation in classification performance. In order to have maximum time resolution the segmentation length was set to 4 s. The second segmentation parameter, the time offset between the segment center and the beginning of the MSE, showed a very sensitive influence on classification performance (Fig. 1). Only in a small interval the algorithms can learn accurately; it is optimal to set the center of the segment 1 s after MSE starts. Also, LogPSD features resulted in lower error rates than DWT features.

For these optimal settings, the comparison of LVQ, GB, and SVM learning algorithms resulted in small differences in mean errors and their standard deviations (Tab. 1). The computational load of SVM was at least 10 times higher than that of GB and this was about the same factor higher than that of LVQ. It must be emphasized, however, that these are results from repeated random subsampling of independent training and test sets that include data from all subjects in both sets. Thus, the methods already have information about each subject during the training. This is not the case for LOSO CV, in which data of one subject is held out of training and used to estimate the mean test error. For 31 subjects, the average size of the training set is $30/31 N = 7648$ and $1/31 N = 255$ for the test set. Since the data of each test person are kept out and the data of all others are used for training, there are 31 different mean test errors (Fig. 2). The results show that there are large differences in learning performance. Data from most subjects were classified accurately by classifiers who learned from data from all other subjects. However, if only the results of the best classifier (SVM) are considered, it turns out that five subjects are classified with errors higher than 10 %. That is, during training the methods did not obtain enough information from data of the other 30 subjects to correctly classify data of the subject held out. Comparison of learning methods shows that

in LOSO CV the SVM can almost always classify better than GB and LVQ, often much better.

5 Conclusions

The presented investigation shows that classification of short-term EEG segments to behavioral characteristics is possible at low error rates, if the characteristics relate to a change in brain state, i.e. the microsleep. It has been demonstrated that only in a short time interval around the onset of MSE an accurate learnability is given. For segments that are a few seconds before or after MSE onset, such low errors cannot be achieved, because obviously no specific brain state can be defined here, but the normal multi-process mode of the awake brain, which is limited, however, by high drowsiness.

Direct estimation of LogPSD by the modified periodogram led to higher accuracies than other estimation methods, which provide lower variance. This suggests that the trade-off between bias and variance must be chosen differently for machine learning methods. They obviously benefit from PSD estimation methods with lower bias at higher variance. It remains a future challenge to find a suitable feature extraction from DWT coefficient series. Power and entropy measures were of little use in achieving low classification errors in the following step.

LOSO CV simulates the case that in the future, in a dataset where learning methods had already been learned in the past, another set of examples of a new subject will be added. It was found that data of a few subjects could not be well explained by data of the other 30 subjects. In order to explain this result, the raw recordings and the distribution of the extracted features were inspected. No easily identifiable cause was found. One subject (#30) had unusually strong and long-lasting blink activity in periods of high drowsiness. This activity superimposed all channels and may have led to a considerable bias in estimated spectral features. Others were found to have poor signal quality, probably due to poor electrode contact resistance in combination with low amplitude EEG.

It might also be that the individually diverse EEG characteristics are a further explanation. This has already been reported from studies on features extracted from 30 s EEG segments during sleep. In studies with monozygotic and dizygotic twins, evidence was found of high individuality and high heritability of EEG features. One group of authors concluded that the EEG could possibly be the most inheritable trait of humans [8]. We are working on carefully processing data from further studies in our laboratory and incorporating them into LOSO

Figure 2: Results of LOSO cross-validation of all subjects. Mean classification errors for three different learning algorithms.

CV analyses in order to ultimately obtain indications for a conclusive explanation.

Author Statement

Authors state no funding involved, as well as no conflict of interest. Informed consent has been obtained from all individuals included in both studies. Ethical approval: The research related to human use complies with all 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.

References

- [1] Golz M, Sommer D. Automatic Recognition of Microsleep Events. *J Biomedizinische Technik* 2004;49,Suppl2:332-3.
- [2] Peiris M, Jones R, Davidson P, Bones P. Detecting behavioral microsleeps from EEG power spectra. *Proc 28th EMBS Conf* 2006:5723-26.
- [3] Golz M, Schenka A, Sommer D, et al. The role of expert evaluation for microsleep detection. *Current Directions in Biomedical Engineering* 2015;1(1):92–5.
- [4] Xu G, Huang JZ. Asymptotic optimality and efficient computation of the leave-subject-out cross-validation. *The Annals of Statistics* 2012;40(6):3003-30.
- [5] Golz M, Sommer D, Trutschel U, et al. Evaluation of fatigue monitoring technologies. *J Somnology* 2010;14(3):187-99.
- [6] Misiti M, Misiti Y, Oppenheim G, Poggi JM. Wavelet toolbox. The MathWorks Inc., Natick (MA), USA; 2017.
- [7] Ke G, Meng Q, Finley T, Wang T, Chen W, et al. LightGBM: A highly efficient gradient boosting decision tree. In *Adv Neural Inform Proc Syst* 2017:3146-54.
- [8] De Gennaro L, Marzano C, Fratello F, et al. The EEG fingerprint of sleep is genetically determined: A twin study. *Annals of Neurology* 2008;64:455-60.