The role of expert evaluation for microsleep detection

Abstract: Recently, it has been shown by overnight driving simulation studies that microsleep density is the only known sleepiness indicator which rapidly increases within a few seconds immediately before sleepiness related crashes. This indicator is based solely on EEG and EOG and subsequent adaptive pattern recognition. Accurate microsleep recognition is very important for the performance of this sleepiness indicator. The question is whether expensive evaluations of microsleep events by a) experts are necessary or b) non-experts provide sufficient evaluations. Based on 11,114 microsleep events in case a) and 12,787 in case b) recognition accuracies were investigated utilizing (i) artificial neural networks and (ii) support-vector machines. Crossvalidated classification accuracies ranged between 92.2% for (i, b) and 99.3% for (ii, a). It is concluded that expert evaluations are very important to provide independent information for detecting microsleep.

Keywords: Sleepiness; Microsleep; Driving Simulation; EEG; EOG; Neural Networks; Support-Vector Machines

DOI: 10.1515/CDBME-2015-0024

1 Introduction

The assessment of driver sleepiness is still a challenging task from the technological as well as scientific point of view. It is very important in the light of risk management, since drowsy driving represents 10 to 30 percent of all crashes [1]. It was estimated that 160,000 injuries and 6,000 deaths each year in the European Union are primarily caused by sleepiness (cf. [1] and references therein).

Using our driving simulator, it has been demonstrated that immediately before crashes several sleepiness indicators do not change, with the exception of EEG / EOG microsleep density [2]. Only microsleep density zooms up within a few seconds. This indicator is defined as the percentage of microsleep patterns found in EEG and EOG within an accumulation interval of 2 min in length. Machine learning algorithms are used for the detection of microsleep events (MSE). During supervised learning, these algorithms require EEG / EOG as input data and information on the desired output as well. The latter is a binary variable and consists of either MSE or counterexamples (Non-MSE). MSE can be obtained by the following determination methodologies. Non-MSE are periods in time where drivers are drowsy but still able to drive.

The determination of MSE includes some trade-offs. One way is to define them only by EEG criteria [3]. It is assumed that spectral deceleration of the EEG is typical for MSE. Commonly, an EEG expert searches for theta activity which has replaced waking alpha background. However, there are a number of decelerations without concomitant impairment in driving ability [4]. Furthermore, several authors observed that alpha not theta activity plays a prominent role in driver sleepiness (e.g. [5]).

A second type of MSE determination is based on behaviour and assumes that lapses in motor response within continuous steering tasks are typical for MSE [6]. The problem here is that other causes like e.g. cognitive distraction or lack of concentration not related to sleepiness may be present. Advantageously, all dangerous performance lapses are included in analyses, independent of their inner causes. Recently, the same group [6] presented a determination combined with the third type [7].

In the third type of MSE determination a trained person performs subjective video ratings of driver’s behaviour, especially based on eye and head movements. If prolonged eye lid closures, rolling eye movements, slow head movements, head nods, or similar typical signs of MSE exhibit an event is annotated manually. It must be emphasized that a precise definition of MSE cannot be made due to the inter-individual differences in behaviour; it remains a subjective decision.

Advantageously, during evaluation the observer traces the temporal development of driver’s behaviour and is prepared to expect occurrence of MSE.

In the following, the last kind of MSE determination is regarded. For this purpose, highly motivated and trained observers are requested. It is asked whether longstanding experts with deep knowledge and wide experiences are es-
2 Material and methods

2.1 Driving simulation studies

Detailed information on this topic has been published earlier [5]. In short, 16 young adults (mean age 24.4 ± 3.1 years) were randomly selected out of 133 interested volunteers. Two experimental nights per subject were conducted. During three days before experiments it was checked by actimetry whether the scheduled sleep wake restrictions were fulfilled (wake up time: 6:00-8:00, time to bed: 22:00-1:00). Furthermore, it was checked whether during the day before experiments no short sleep (naps) happened. 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).

Behavioural MSE were evaluated by a) one expert with longstanding knowledge and experiences in the field, and by b) 9 non-experts with some knowledge and experiences lasting not longer than 9 months. Based on visual evaluations of video material, of lane deviation time series and of EOG, their task was to search for critical events and to assign each event to one of 6 severity levels (1 = vague or very short signs of MSE lasting not longer than 0.3 sec.; 2 = short MSE lasting between 0.3 and 1.5 sec., 3 – 6 = extended MSE with duration of at least 2.5, 3.5, 4.5, and 5.5 sec., respectively). Lane deviation is an output variable of the driving simulator and is the lateral position of the car with respect to the centre of the lane.

In addition, the starting time of each MSE was determined as accurately as possible. For supervised learning methods counterexamples (Non-MSE) are needed, i.e. periods in time where the driver is drowsy but still able to keep the car in lane. Logarithmic power spectral densities averaged in narrow bands (0.5 <f <23.5 Hz, Δf = 2 Hz) were estimated using modified periodogram. Alternatively, Welch’s and multi-taper method were applied, but resulted in slightly lower classification accuracies. Delay-vector variances were further features extracted from all signals [9]. After that, feature vectors were processed by LVQ (learning vector quantization) and SVM (support-vector machines) in order to assign them to class labels (MSE or Non-MSE). This classification step included machine learning, where internal parameters of LVQ and SVM are optimized. After finishing learning, cross validation analysis was performed; outputs were mean and standard deviation of classification accuracies for training as well as validation data. Both data sets were constituted by multiple, random subsampling. Parameters of all processing steps were optimized such that training set accuracies were maximal (see Figure 1). Validation set accuracy is the main outcome; it is an empirical estimate of the true classification performance.

2.2 Pattern recognition

Data analysis was based on clearly visible MSE (level 2 – 6). Pre-processing included signal segmentation for each MSE and for the same amount of Non-MSE. As mentioned above, Non-MSE are periods in time where subjects were drowsy, but were able to keep the car in lane. Logarithmic power spectral densities averaged in narrow bands (0.5 <f <23.5 Hz, Δf = 2 Hz) were estimated using modified periodogram. Alternatively, Welch’s and multi-taper method were applied, but resulted in slightly lower classification accuracies. Delay-vector variances were further features extracted from all signals [9]. After that, feature vectors were processed by LVQ (learning vector quantization) and SVM (support-vector machines) in order to assign them to class labels (MSE or Non-MSE). This classification step included machine learning, where internal parameters of LVQ and SVM are optimized. After finishing learning, cross validation analysis was performed; outputs were mean and standard deviation of classification accuracies for training as well as validation data. Both data sets were constituted by multiple, random subsampling. Parameters of all processing steps were optimized such that training set accuracies were maximal (see Figure 1). Validation set accuracy is the main outcome; it is an empirical estimate of the true classification performance.

Figure 1: Pattern recognition consisted of four different stages. Parameters of the first three stages were optimized in order to gain maximal accuracy (ACC) on input data (training set). Input data were signal features of EEG / EOG as well as class labels, i.e. microsleep events (MSE) and non-microsleep (Non-MSE), for further explanations see text.

Table 1: Number of MSE for different severity levels (see text) depending on visual evaluation by a) expert and b) non-expert.

<table>
<thead>
<tr>
<th>Severity level</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>a) Expert evaluation</td>
<td>13,095</td>
<td>6,182</td>
<td>2,806</td>
<td>1,231</td>
<td>573</td>
<td>322</td>
</tr>
<tr>
<td>b) Non-expert evaluation</td>
<td>4,830</td>
<td>5,846</td>
<td>3,861</td>
<td>1,596</td>
<td>880</td>
<td>604</td>
</tr>
</tbody>
</table>
3 Results

Recordings of 32 overnight experiments (16 subjects, 2 nights each) were evaluated. Each night consisted of 8 driving sessions of 40 min duration each, resulting in 10,240 min total driving time. Each of 9 different non-experts evaluated driver behaviour for some driving sessions. There was no fixed allocation; non-experts evaluated between 10 and 30 driving sessions. In contrast, the expert evaluated all 256 driving sessions (32 nights, 8 sessions each). A lot of differences occurred between both evaluations (Table 1). The sum of all events including vague events (level 1) was a) 24,209 and b) 17,617. Clearly visible MSE (level 2 – 6) added up to a) 11,114 and b) 12,787.

Expert and non-expert evaluations resulted in different annotations for the following three reasons:
1. Different severity levels of MSE,
2. Different adjustment of the starting times of MSE,
3. Different adjustment of the starting times of Non-MSE.

In consequence, two different sets of MSE and Non-MSE were used in further signal processing and subsequent classification. As a result of many parameter optimizations, mean and standard deviation of classification accuracy were computed. Due to the paradigm of cross validation, classification accuracies were estimated on training data (Table 2) as well as on test data (Table 3). The first is an inspection of the adaptivity of the classifier and indicates numerical problems if existing. The second is an inspection of the ability of the pattern recognition methodology to generalize and is an estimation of the true classification accuracy. It gives an impression in which range accuracies for future examples should lie provided that a representative sample is given.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>LVQ</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>a) Expert</td>
<td>99.45 ±0.17 %</td>
<td>99.92 ±0.00 %</td>
</tr>
<tr>
<td>b) Non-expert</td>
<td>94.81 ±0.35 %</td>
<td>99.09 ±0.00 %</td>
</tr>
</tbody>
</table>

4 Conclusion

Results offer remarkable differences between a) expert and b) non-expert evaluation. Non-experts tended to overestimate severity level of MSE and evaluated much more events on levels 3 – 6 than the expert. On the other hand, a lot of vague events were not noticed by non-experts. Vague events were not included within analysis presented, but will be subject of future investigations. Another key difference between a) and b) lies in determining MSE starting time. Several larger errors of the non-experts were found by the expert. They were not comprehensible and are presumably due to limited endurance and low motivation. Visual evaluations are common in clinical practise and in life science research. They are known to be tedious and demotivating.

The expert himself reported that he has continuously extended his intrinsic criteria for his decisions, because behaviour during extreme fatigue was relatively complex and differed largely between subjects. Therefore, it is likely that he would classify some events different if he would again evaluate the recordings. It is an open question how large the resulting alterations would be and if machine learning algorithm would even better work or not.

Support-vector machines outperformed learning vector quantization. This result shows that strict mathematical concepts, like large margin optimization and non-linear mapping to higher-dimensional spaces via kernel functions, are superior to stochastic optimization concepts with less mathematical rigidity. The machine learning process of support-vector machines has much more computational load as of learning vector quantisation; the difference was in the region of up to 10^6. For very large data sets this might pose a problem.

Further relevant improvements in classification accuracies are hardly to obtain, especially for support-vector machines, because 99.3 % is already a very high value. But it must be emphasized that data of all drivers were part of training as well as validation data. The learning classifiers were adapted to data of all drivers and it was possible to generalize (Table 3), i.e. accuracies of unlearned data.
were almost as large as of learned data. Further investigations should include subject hold-out validation, which means that training data consists of MSE of all but one subject. Validation data consist only of data of the one subject which is hold-out from the training set. If this procedure was repeated for every subject, mean and standard deviation would give estimation on how accurate MSE can be recognized for drivers which may be included in future investigations with the same methodology.

In summary, it can be stated that non-expert evaluation of MSE is sufficient for high recognition success, but only expert evaluation paves the way for highly accurate MSE detection. This way, a laboratory reference standard of driver sleepiness has been established which can be used to evaluate devices for fatigue monitoring [8].

Another important application for applied research and maybe for future warning systems might be an on-line closed-loop MSE detection and mitigation system based on dry electrode EEG [10]. If MSE were detected then immediate auditory warnings were presented to the driver. It has been demonstrated that the effectiveness of arousing auditory warnings might depend on EEG spectra features [10]. Effective warnings led to improved driver’s response times to subsequent lane departure events. Furthermore, it has been demonstrated that for upcoming MSE the EEG power spectral densities immediately changed and that through warnings they came back to signatures which are typical for the alert state without bouncing back to the drowsy level. In future, this might lead to real-life applications of the dry and wireless EEG technology based on smartphones as mobile signal processing platform.

**Funding:** This study was funded by the Federal Ministry of Education and Research within the research program “Research at University of Applied Sciences together with Enterprises” under the project 176X08.

**Author’s Statement**

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

**References**


