Driver sleepiness detection with deep neural networks using electrophysiological data

Martin Hultman, Ida Johansson, Frida Lindqvist and Christer Ahlström

Driver sleepiness detection with deep neural networks using electrophysiological data

Martin Hultman[1], ORCID: 0000-0001-5912-3281

Ida Johansson[1], ORCID: 0000-0002-1404-9177

Frida Lindqvist[1], ORCID: 0000-0001-7288-2581

Christer Ahlström[1,2,*], ORCID: 0000-0003-4134-0303

[1]Department of Biomedical Engineering, Linköping University, Linköping, Sweden

[2]Swedish National Road and Transport Research Institute (VTI), Linköping, Sweden

[*]Corresponding author: christer.ahlstrom@vti.se

Abstract

Objective: The objective of this paper is to present a driver sleepiness detection model based on electrophysiological data and a neural network consisting of Convolutional Neural Networks and a Long Short-Term Memory architecture.

Approach: The model was developed and evaluated on data from 12 different experiments with 269 drivers and 1187 driving sessions during daytime (low sleepiness condition) and night-time (high sleepiness condition), collected during naturalistic driving conditions on real roads in Sweden or in an advanced moving-base driving simulator. Electrooculographic and electroencephalographic time series data, split up in 16634 2.5-minute data segments was used as input to the deep neural network. This probably constitutes the largest labelled driver sleepiness dataset in the world. The model outputs a binary decision as alert (defined as ≤6 on the Karolinska Sleepiness Scale, KSS) or sleepy (KSS≥8) or a regression output corresponding to KSS ϵ [1-5,6,7,8,9].

Main results: The subject-independent mean absolute error (MAE) was 0.78. Binary classification accuracy for the regression model was 82.6% as compared to 82.0% for a model that was trained specifically for the binary classification task. Data from the eyes were more informative than data from the brain. A combined input improved performance for some models, but the gain was very limited.

Significance: Improved classification results were achieved with the regression model compared to the classification model. This suggests that the implicit order of the KSS ratings, i.e. the progression from alert to sleepy, provides important information for robust modelling of driver sleepiness, and that class labels should not simply be aggregated into an alert and a sleepy class. Furthermore, the model consistently showed better results than a model trained on manually extracted features based on expert knowledge, indicating that the model can detect sleepiness that is not covered by traditional algorithms.

Keywords: Sleepiness detection, driving, EEG, EOG, deep learning

1 Introduction

About 1.35 million people are killed in road crashes every year (World Health Organization, 2018), and it is estimated that driver fatigue, including sleepiness, contributes to about 20 % of these deaths (Åkerstedt, 2000, Connor et al., 2002). Enforcement officers report difficulties in identifying driver fatigue (Radun et al., 2013) and there is a high degree of underreporting (Phillips and Sagberg, 2013). Yet, crash rates where fatigue was reported as a contributing factor are significantly higher than baseline crash rates, and drivers reporting sleepiness at the wheel have more than a two-fold increase of being involved in a crash (Bioulac et al., 2017).

Fatigue may result from various causes such as physical exertion, lack of physical activity, insufficient sleep, boredom, worry or overwork. In this paper we focus on fatigue due to sleepiness as an effect of insufficient sleep in combination with night-time driving in the window of circadian low. Sleep-related forms of fatigue are characterised by a slowing of visual processing, loss of selective attention, distractor inhibition, reduced peripheral processing capacity and wake state instability (Chee, 2015, Krause et al., 2017), all of which are detrimental for safe driving. Driver fatigue detection systems can here be used to convince fatigued drivers to pull over for a rest or nap, or to detect when a driver is unfit to take back control from an automated vehicle.

There has been considerable development in the area of driver fatigue detection (see for example the following reviews: Balandong et al., 2018, Liu et al., 2009, Ramzan et al., 2019, Sikander and Anwar, 2018, Golz et al., 2010, Chowdhury et al., 2018, Sahayadhas et al., 2012). Fatigue detection systems are typically based on vehicular, behavioural, or physiological information. Vehicle-based measures, such as steering wheel activity and lane positioning, are already available in today's vehicles and have the advantage of being nonobtrusive. However, the warnings are often inaccurate and unreliable (Sahayadhas et al., 2012). Behavioural measures, such as gaze behaviour and eye closures, can be camera-based and are thus nonintrusive, but the extraction of eye features from video data is still difficult, especially in direct sunlight or quickly changing light conditions (Fernández et al., 2016). Physiological measures have been found to be more reliable (Sahayadhas et al., 2012), but they are as of yet very obtrusive. This paper focus on sleepiness detection based on physiological measures.

When designing fatigue classification systems based on physiological data, the main sources of information are electroencephalography (EEG) to measure brain activity, electrooculography (EOG) to measure eye and blink behaviour, and electrocardiography (ECG) or photoplethysmography (PPG) to measure heart rate and heart rate variability (HRV). EEG data are typically quantified using the power in the theta (4–7 Hz), alpha (8–15 Hz) and beta (16–31 Hz) frequency bands, where increased theta power reflects sleep need (Shahid et al., 2010) and increased alpha power indicates driver sleepiness (Kecklund and Åkerstedt, 1993, Simon et al., 2011). EOG data are typically quantified in terms of blink durations or eyelid opening/closing velocities (eg. Schleicher et al., 2008, Åkerstedt et al., 2005), whereas cardiorespiratory function is measured with different HRV metrics (van den Berg et al., 2005, Apparies et al., 1998, Tran et al., 2009, Patel et al., 2011, Yang et al., 2010, Vicente et al., 2016). On the negative side, EEG-based measures suffer from noise in naturalistic settings, large inter-individual variability, and the fact that some individuals do not respond despite being clearly sleepy (Sparrow et al., 2018, Sandberg et al., 2011a, Åkerstedt et al., 2010). Cardiorespiratory metrics often show ambiguous results since they not only by modulated by sleepiness and fatigue but also by several other time-varying inter- and intra-individual factors such as age, gender, posture, distress, boredom and relaxation (Persson et al., 2019). Finally, EOG metrics suffer from individual differences and small effect sizes (e.g. Caffier et al., 2003, Campagne et al., 2005, Ingre et al., 2006b, Schleicher et al., 2008, Papadelis et al., 2007).

Numerous signal analysis methods have been used to extract fatigue indictors, or features, based on domain knowledge. These include phase synchronization (Kong et al., 2017), spectral analysis (Fu et al., 2016, Golz et al., 2007, Hu et al., 2013, Li et al., 2015, Picot et al., 2012), joint time/frequency and wavelet analysis (Golz et al., 2016, Khushaba et al., 2011, Liu et al., 2010), and nonlinear approaches such as fractal dimensions and different entropies (Golz et al., 2007, Mu et al., 2017, Papadelis et al., 2007). These features are then often used as input to various machine learning algorithms, such as neural networks (Lin et al., 2012, Patel et al., 2011, de Naurois et al., 2018), support vector machines (Golz et al., 2007, Golz et al., 2016, Hu et al., 2013, Khushaba et al., 2011, Li and Chung, 2015, Li et al., 2015, Mårtensson et al., 2019), and hidden Markov models (Fu et al., 2016, Liu et al., 2010, Yang et al., 2010).

Another approach is to include the feature extraction step in the machine learning model using deep learning. Detection of different stages of fatigue using deep learning is a relatively new and unexplored area. However, deep learning has successfully been used for automatic classification of physiological signals in other domains. Roy et al. (2019) reviewed 156 papers on deep learning-based EEG analysis. In their review, there were no papers on sleepiness or fatigue scoring, but 15 that investigated the related topic of sleep staging. Most of the reviewed work used pre-processing steps such as filtering of the EEG data and a few papers used artifact handling. Roughly half of the reviewed papers used frequency domain features as raw EEG data features while other papers used the raw EEG time series. The selection of network architectures was reported to vary over the years but in total almost half (41%) used Convolutional Neural Networks (CNN) and a smaller amount (14 %) used Recurrent Neural Networks (RNN) or Autoencoders (13 %). Hybrid approaches combining CNNs and RNNs were less common (7 %) but showed an increasing trend.

A few recent papers have presented results on driver fatigue classification based on deep learning and EEG-data. Chai et al. (2017) used sparse-deep belief networks and reached an accuracy of 93.1% on a dataset with 43 participants driving on a monotonous road in a driving simulator. Zeng et al. (2018) used a combination of CNN and deep residual learning and reached an accuracy of 92.7% (subject-dependent classification) and 84.4% (subject-independent classification) on a dataset with 10 participants driving on a monotonous road in a driving simulator. Ma et al. (2019) used a principal component analysis network and reached an accuracy of 95% on a dataset with 5 participants, also in a driving simulator setting. Song et al. (2019) used a CNN and reached an accuracy of 75.9 % on a dataset with 36 participants. Here, the participants were instructed to perform a series of facial expressions such as yawning, slow blink rate and falling asleep while driving in a driving simulator. A limitation that these studies have in common is that the work was carried out on small datasets with low ecological validity. Further, fatigue was invoked either as task related underload or by asking the participants to act sleepy. This is in stark contrast with work on sleep stage classification based on deep learning, where large clinical databases with thousands of participants are used.

The work presented in this paper is inspired by recent work on automatic sleep staging using deep recurrent and convolutional neural networks (Biswal et al., 2017, Biswal et al., 2018, Stephansen et al., 2018). The aim is to explore the potential of deep learning to classify different levels of driver sleepiness based on electrophysiological data. A secondary goal is to compare the classification performance of deep features extracted by the deep learning model versus manually defined features based on domain knowledge. A third goal is to investigate if EEG, EOG or a combination of EOG and EEG is most suitable for driver sleepiness classification.

2 Methods

Two methodological approaches have been used for driver sleepiness classification in this paper. A classic machine learning framework with feature extraction based on domain knowledge combined with shallow learning and a deep learning framework where relevant features are automatically extracted from the electrophysiological data. Common for both approaches were a pre-processing stage where EOG and EEG data were filtered, normalized, and divided into 2.5-minute segments. Pre-processing were carried out in MATLAB 2019a (Mathworks Inc., Natick, MA, USA) whereas the neural networks were implemented using Keras deep learning library in Python 3.7 using the TensorFlow backend.

2.1 Sleepiness database

Datasets from 12 separate driver sleepiness experiments were combined in this paper. Five of the experiments were run on real roads, either on a rural road or on a highway outside Linköping, Sweden. The remaining seven experiments were run in a high-fidelity moving-base driving simulator[1]. The cars used in the on-road experiments were equipped with dual control to allow a safety driver to intervene if needed. Permission to conduct driving sessions with sleep deprived drivers on public roads was given by the Swedish government (N2007/5326/TR). All experiments were approved by the Swedish Ethical Review Authority, see the referenced papers in Table 1 for more detailed information.

The participants were recruited by random selection from the Swedish register of vehicle owners. All drivers were prepared in a similar way in all experiments. Before arrival, the participants were requested to avoid alcohol for 72 hours and to abstain from nicotine and caffeine for 3h before driving. All participants reported that they were healthy with good to excellent sleep quality. In 11 of 12 experiments, the 225 drivers drove in at least one alert condition during daytime and one sleep deprived condition during night-time after being awake since early morning. In the 12th experiment, the 44 drivers only drove in a sleep deprived state in the early morning hours after a night shift. The duration of the driving sessions ranged from 30 to 90 minutes.

Electrophysiological data (EEG and EOG) were recorded with a bio-amplifier (Vitaport 2 or 3, Temec Instruments BV, the Netherlands or g.HIamp, g.tec Medical Engineering GmbH, Austria). The electrodes used for the EOG (measured vertically and horizontally across the eyes) were of the disposable Ag/AgCl type. The EEG was measured via three bipolar derivations positioned at Fz-A1, Cz-A2 and Oz-Pz using silver plated non-disposable electrodes.

The Karolinska Sleepiness Scale (KSS) was used to acquire self-reported sleepiness every fifth minute during the drives. KSS has nine anchored levels (Åkerstedt and Gillberg, 1990): 1 - extremely alert, 2 - very alert, 3 - alert, 4 - rather alert, 5 - neither alert nor sleepy, 6 - some signs of sleepiness, 7 - sleepy, but no effort to keep alert, 8 - sleepy, some effort to keep alert, and 9 - very sleepy, great effort to keep alert, fighting sleep. The reported value corresponds to the average feeling during the past 5 minutes. The KSS values were used as the target values when training the machine learning algorithms.

Table 1: Summary of datasets used to train and test the developed classifiers.

Dataset	Driving environment	Number of drivers	Number of sessions	Reference
1	Road	24	3	Anund et al. (2013)
2	Road	43	3	Hallvig et al. (2014)
3	Road	18	2	Silveira et al. (2019)
4	Sim	14	6	Åkerstedt et al. (2010)
5	Road	18	5	Sandberg et al. (2011a)
6	Road	24	2	Schwarz et al. (2012)
7	Sim	12	4	Radun et al. (2014)
8	Sim	12	3	Leandersson Olsson (2012)

[1] https://www.vti.se/en/research-areas/vtis-driving-simulators/

9	Sim	30	18	Barua et al. (2019)
10	Sim	10	2	Ingre et al. (2006a)
11	Sim	20	2	Bekiaris et al. (2005)
12	Sim	44	1	Anund et al. (2008)

2.2 Preprocessing

The EOG and EEG time series were split into 2.5-minute segments. Since the KSS-ratings were given every 5th minute, this provided two segments per KSS-rating. The 2.5-minute segment duration was chosen as a compromise between having enough raw data to calculate the features, and short enough to give a reasonably fast sleepiness detection.

The EOG signals were lowpass filtered with a 5th order Butterworth filter at 11.52 Hz. Baseline drift and saturation due to motion artifacts was removed by subtracting a piecewise linear function using breakpoints located at large absolute derivatives >50mV/s and at the borders of saturated segments. Each 2.5-minute EOG-segment was then normalised by subtracting the mean value and division by the median blink amplitude. The blink amplitudes were extracted according to Jammes et al. (2008). Segments with a maximum amplitude below 5 µV were removed since such low amplitudes is an indication of a detached electrode.

EEG data were bandpass filtered with a 5th order Butterworth filter with cut-off frequencies at 0.3 Hz and 40 Hz. EEG-segments with amplitudes above 200 µV were considered as artifacts and removed from further analyses, as were segments with maximum amplitudes below 5 µV. This is a very rudimentary artifact handling procedure, and the aim was simply to remove the worst outliers from the data. Each EEG-segment was normalized using the 1st and 99th percentiles of the signal to robustly squeeze the values into a range close to [-1,1]. All filtering was done with zero-phase forward and reverse digital IIR filtering.

Sleepiness detection was performed using both binary classification to classify alert vs. sleepy (where alert was defined as KSS ≤ 6 and sleepy as KSS ≥ 8), and regression to estimate the reported KSS score on a finer level (KSS ϵ [1-5,6,7,8,9]). To obtain a clear separation of the two binary classes, segments with KSS = 7 were discarded as outlined by Sandberg et al. (2011b). The class definitions are justified by the observation that hardly any line crossings, i.e. when the vehicle is about to veer out of lane, occur at KSS ≤ 6, whereas a markedly increased frequency of unintentional lane deviations occurs at KSS ≥ 8 (Hallvig et al., 2014, Åkerstedt et al., 2014). In the regression model, KSS 1–5 were merged into one alert class since these ratings all represent different levels of “non-sleepiness”. These class definitions resulted in the label distribution presented in Figure 1.

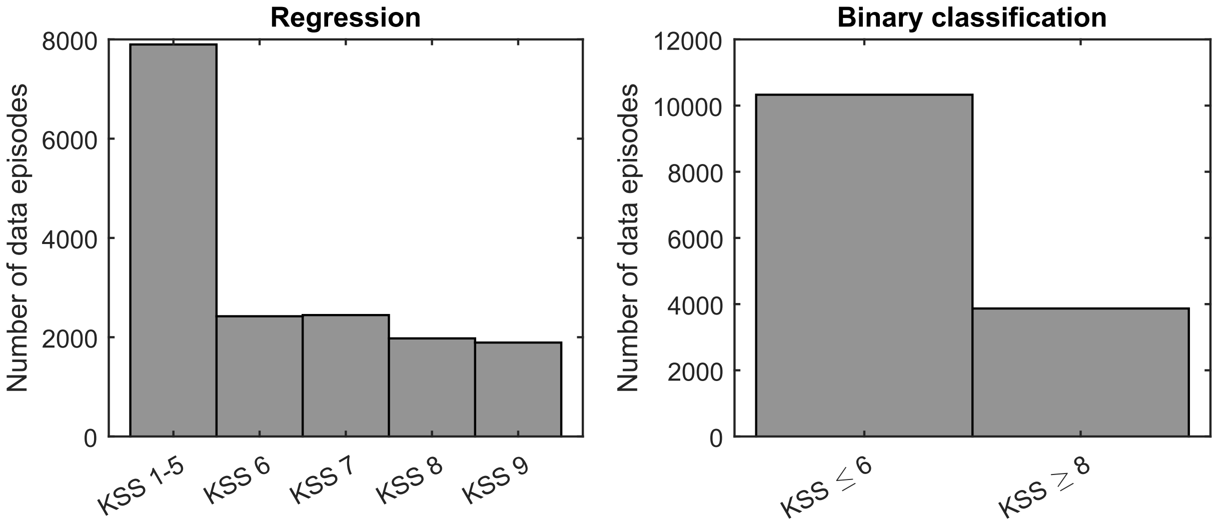


Figure 1: Label distribution in the case of regression and binary classification, respectively.

2.3 Classification with manually extracted features

Blink behaviour features were extracted from the vertical EOG signal with an automatic blink detection algorithm based on derivatives and thresholding (Jammes et al., 2008). For each eye blink, the blink duration, eyelid closure speed, peak closing speed, eyelid opening speed, peak opening speed, delay of eyelid reopening, closing time and opening time were extracted. To reduce problems with concurrence of eye movements and blinks, the blink duration was calculated at half the amplitude of the upswing and the downswing of each blink and defined as the time elapsed between the two. The blink parameters were extracted in each 2.5-minute segment and summarized using the 1st percentile, 25th percentile, 50th percentile, 75th percentile and 99th percentile of each blink parameter. The resulting shape of the input data for the classifier was (N, 45), where N is the number of segments, and the 45 features are the 5 percentiles for each of the 9 blink parameters.

Brain activity features were extracted using a frequency representation of the EEG data from the Oz-Pz channel. The power spectral density was estimated via Welch's method using a window size of 1 second, an overlap of 0.5 seconds, and 31 frequency bins ranging from 0-30 Hz. The resulting shape of the input data to the classifier was (N, 31), where N is the number of segments.

A shallow feedforward network architecture was used, consisting of a fully connected layer using ReLU activation functions and 64 nodes, followed by the output layer with one node using linear activation in the case of regression and two nodes with softmax activation in the case of binary classification. In the regression case, the order of the ordinal KSS scale was conveniently considered by the imposed regression structure. Three input modalities were used; only EOG features, only EEG features, as well as EOG and EEG features combined. In the third case the features from EOG and EEG were concatenated to a single channel with 76 features. The model is described in detail in Table 2.

Table 2: Detailed summary of the fully connected network architecture. Note that the size of the input and output layers depend on the data and task (regression or binary classification), respectively. The batch size is not included in the output shape.

Layer		Parameters	Output shape
0	Input	EOG input EEG input EOG+EEG input	(45) (31) (76)
1	Dense	64 output features	(64)

2	ReLU		(64)
3	Dense	1 output feature (regression) 2 output features (binary classification)	(1) (2)
4	Output activation	Linear (regression) Softmax (binary classification)	(1) (2)

2.4 Deep learning classification

A network combining a CNN architecture and a recurrent Long Short-Term Memory (LSTM) architecture was set up based on earlier work on sleep staging (Biswal et al., 2017). The CNN network aimed to find temporal patterns in the input data while the LSTM network was intended to account for long-term dependencies in the automatically extracted features. The proposed network design is visualized in Figure 2.

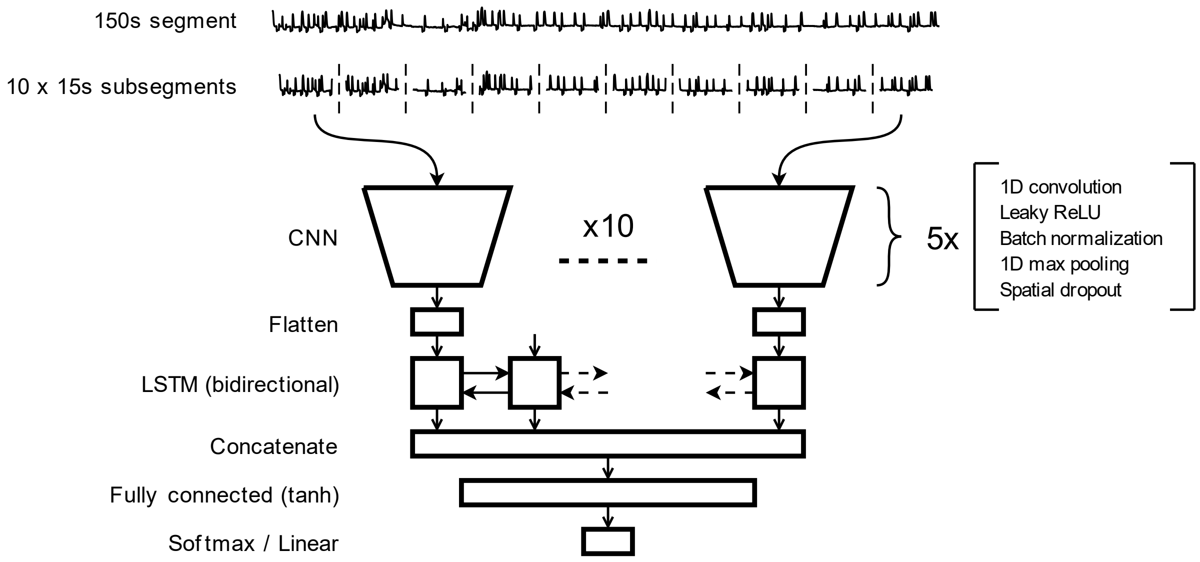


Figure 2: Schematic overview of the CNN-LSTM network architecture. For more details, see Table 3.

The EEG and vertical EOG data that were used as input to the CNN-LSTM network were down-sampled to 64 Hz. To allow the use of an LSTM, each of the N 2.5-minute segments was split into T=10 subsegments of length L=960 sample points (64Hz × 150s / 10). The data were then structured in an input tensor of size (N, T, L, C), where N is the number of segments, and C is the number of channels (C=1 for EOG, C=3 for EEG, and C=4 when both EEG and EOG was used as input).

The CNNs of the network consisted of 5 convolution blocks, each comprising a 1D convolutional layer (filter size 5), leaky ReLU activation ($\alpha = 0.1$), max-pooling (size 3), batch normalization, and 1D spatial dropout (20%). The output of the CNNs was flattened and the extracted features for each subsegment fed to a bidirectional LSTM with 10 cells. The output from each cell was concatenated and fed to a fully connected layer with tanh activation function followed by a dropout layer (20%). The output layer was designed in the same way as for the shallow network outlined above, with a 1-node linear activation for the regression task and a 2-node softmax activation for the binary classification. Table 3 summarizes the network architecture in more detail.

The CNNs were implemented using the TimeDistributed wrapper in the Keras deep learning library. This enables weight sharing across the subsegment-dimension of the data, i.e. the 10 CNNs are trained as a single entity and the feature extraction is therefore identical for each subsegment.

Table 3: Detailed summary of the CNN-LSTM network architecture. Note that the size of the input and output layers depend on the data and task (regression or binary classification), respectively. The batch size is not included in the output shape.

Layer			Parameters	Output shape
	0	Input	1 input channel (EOG) 3 input channels (EEG) 4 input channels (EOG+EEG)	(10,960,1) (10,960,3) (10,960,4)
Time Distributed	1	Conv1D	32 filters, kernel size 5	(10,960,32)
	2	Leaky ReLU	α = 0.1	(10,960,32)
	3	Batch norm		(10,960,32)
	4	Max pooling	Kernel size 3	(10,320,32)
	5	Spatial dropout	Probability 20%	(10,320,32)
	6	Conv1D	32 filters, kernel size 5	(10,320,32)
	7	Leaky ReLU	α = 0.1	(10,320,32)
	8	Batch norm		(10,320,32)
	9	Max pooling	Kernel size 3	(10,107,32)
	10	Spatial dropout	Probability 20%	(10,107,32)
	11	Conv1D	32 filters, kernel size 5	(10,107,32)
	12	Leaky ReLU	α = 0.1	(10,107,32)
	13	Batch norm		(10,107,32)
	14	Max pooling	Kernel size 3	(10,36,32)
	15	Spatial dropout	Probability 20%	(10,36,32)
	16	Conv1D	64 filters, kernel size 5	(10,36,64)
	17	Leaky ReLU	α = 0.1	(10,36,64)
	18	Batch norm		(10,36,64)
	19	Max pooling	Kernel size 3	(10,12,64)
	20	Spatial dropout	Probability 20%	(10,12,64)
	21	Conv1D	64 filters, kernel size 5	(10,12,64)
	22	Leaky ReLU	α = 0.1	(10,12,64)
	23	Batch norm		(10,12,64)
	24	Max pooling	Kernel size 3	(10,4,64)
	25	Spatial dropout	Probability 20%	(10,4,64)
	26	Flatten		(10,256)
	27	Bidirectional LSTM	24 output features (in both directions)	(10,48)
	28	Flatten		(480)
	29	Dense	128 output features	(128)
	30	Tanh activation		(128)
	31	Dropout	Probability 20%	(128)
	32	Dense	1 output feature (regression) 2 output features (binary classification)	(1) (2)
	33	Output activation	Linear (regression) Softmax (binary classification)	(1) (2)

2.5 Model training and evaluation criteria

The data were partitioned into three datasets: training 56 %, validation 24 %, and test 20 %. The test set consisted of data from the first 20% of the drivers in each experiment in Table 1. This means that the evaluation was subject-independent, i.e. done on data from drivers that were not used in the training process. The remaining 80% of the data were randomly assigned to the training and validation sets.

The loss function used when training the models was binary cross-entropy in case of binary classification and mean absolute error (MAE) in case of regression. All training was performed using the Adam optimizer with an initial learning rate of 0.0004 for CNN-LSTM networks and 0.001 for shallow networks. L_1 (1%) and L_2 (1%) weight-regularization was used for all layers in the CNN-LSTM. Training was carried out using mini-batch gradient descent with a batch size of 256. The number of epochs was set to 300 for CNN-LSTM networks and 1000 for shallow networks. Early stopping was not used, but the learning rate was reduced by a factor 2 after 50 epochs without significant improvement to the validation loss. These parameters were found empirically to give the best and most consistent results. Class weights were used in the training to correct for class imbalances in the data (see Figure 1).

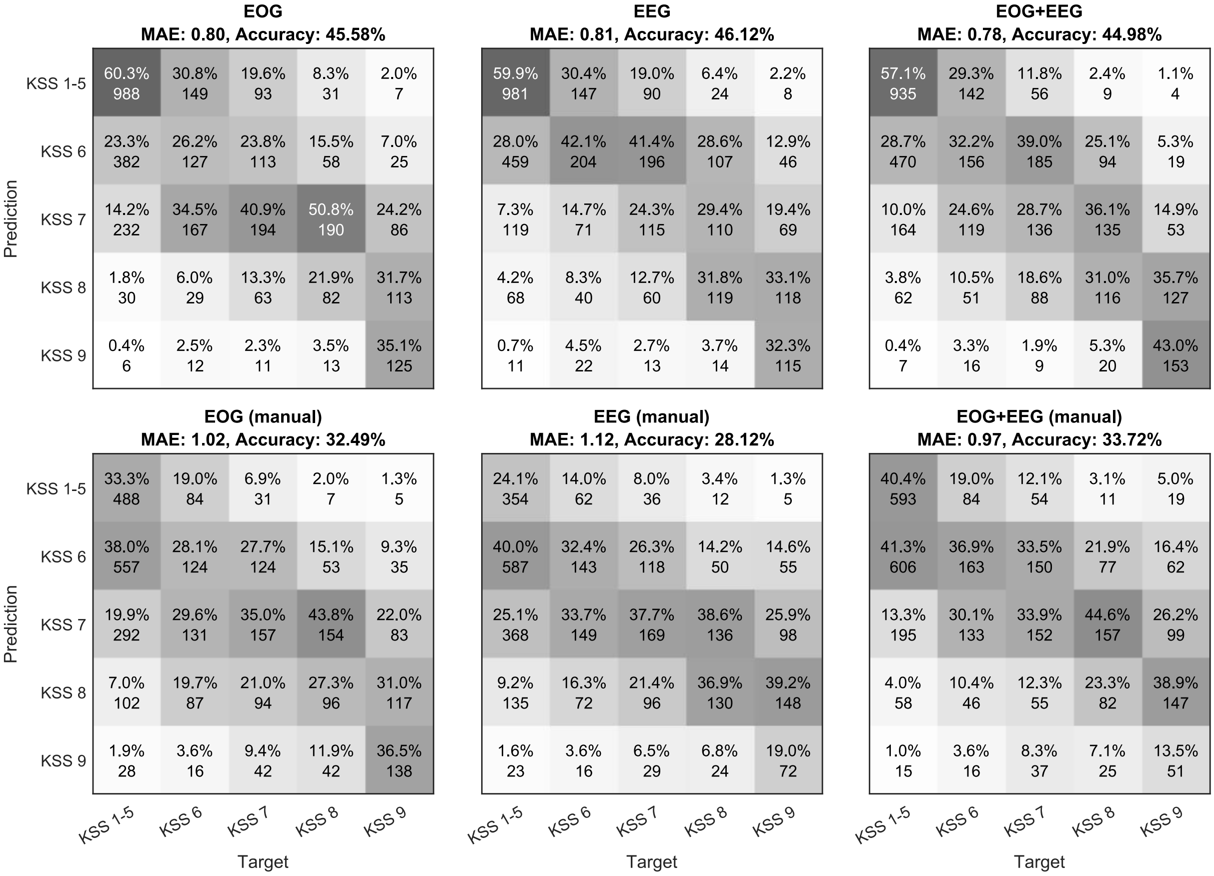
Binary classification performance was evaluated in terms of accuracy, sensitivity, specificity, F1 score and area under receiver operator characteristics curve (AUC). The threshold for calculating the performance metrics was selected to maximize Youden's index, sensitivity + specificity – 1 (Youden, 1950). Regression performance was evaluated in terms of MAE and accuracy, where accuracy was determined after rounding the continuous output to the nearest class label. In addition to this, the regression models were also evaluated as binary classifiers to compare their predictive power with the models specifically trained for binary classification. The threshold used to discretize the continuous regression output into an alert and a sleepy class was, again, determined by maximizing Youden's index. Just as for the binary classification models, this evaluation excluded data with KSS = 7. The main difference between these two binary classification approaches is that the implicit order of the KSS ratings is taken into account in the regression model but not in the model that was trained specifically for the binary classification task.

For each task, 10 separate models were trained from random initial weights. The model with best performance on the validation dataset was evaluated on the test dataset. This is a robust method to select a high-performing model, while ensuring unbiased performance metrics on the test set. For regression networks the validation performance was measured using MAE, and for binary classification AUC was used.

3 Results

Data from the 269 drivers, 1187 driving sessions and 16634 2.5-minute data segments were fed to the different neural network architectures. In general, the deep CNN-LSTM models performed better than the shallow models, and the EOG data scored higher than the EEG data, see Table 4, Table 5, Figure 3 and Figure 4.

Results from the regression analyses are provided in Figure 3. The best performance is achieved when using both EOG and EEG data as input to the CNN-LSTM model. The deep nets have a mean absolute error (MAE) around 0.8 while the shallow nets have an MAE around 1.0, meaning that on average the estimated sleepiness level is about 1 unit off from the reported KSS score. Investigating the confusion matrices in Figure 3, it is however clear that there is great overlap between different KSS ratings despite the low MAE. For example, 17% of the segments where the drivers rated themselves as severely sleepy (KSS≥8) were estimated to be alert (KSS≤6).



EOG
MAE: 0.80, Accuracy: 45.58%

	KSS 1-5	KSS 6	KSS 7	KSS 8	KSS 9
KSS 1-5	60.3% 988	30.8% 149	19.6% 93	8.3% 31	2.0% 7
KSS 6	23.3% 382	26.2% 127	23.8% 113	15.5% 58	7.0% 25
KSS 7	14.2% 232	34.5% 167	40.9% 194	50.8% 190	24.2% 86
KSS 8	1.8% 30	6.0% 29	13.3% 63	21.9% 82	31.7% 113
KSS 9	0.4% 6	2.5% 12	2.3% 11	3.5% 13	35.1% 125

EEG
MAE: 0.81, Accuracy: 46.12%

	KSS 1-5	KSS 6	KSS 7	KSS 8	KSS 9
KSS 1-5	59.9% 981	30.4% 147	19.0% 90	6.4% 24	2.2% 8
KSS 6	28.0% 459	42.1% 204	41.4% 196	28.6% 107	12.9% 46
KSS 7	7.3% 119	14.7% 71	24.3% 115	29.4% 110	19.4% 69
KSS 8	4.2% 68	8.3% 40	12.7% 60	31.8% 119	33.1% 118
KSS 9	0.7% 11	4.5% 22	2.7% 13	3.7% 14	32.3% 115

EOG+EEG
MAE: 0.78, Accuracy: 44.98%

	KSS 1-5	KSS 6	KSS 7	KSS 8	KSS 9
KSS 1-5	57.1% 935	29.3% 142	11.8% 56	2.4% 9	1.1% 4
KSS 6	28.7% 470	32.2% 156	39.0% 185	25.1% 94	5.3% 19
KSS 7	10.0% 164	24.6% 119	28.7% 136	36.1% 135	14.9% 53
KSS 8	3.8% 62	10.5% 51	18.6% 88	31.0% 116	35.7% 127
KSS 9	0.4% 7	3.3% 16	1.9% 9	5.3% 20	43.0% 153

EOG (manual)
MAE: 1.02, Accuracy: 32.49%

	KSS 1-5	KSS 6	KSS 7	KSS 8	KSS 9
KSS 1-5	33.3% 488	19.0% 84	6.9% 31	2.0% 7	1.3% 5
KSS 6	38.0% 557	28.1% 124	27.7% 124	15.1% 53	9.3% 35
KSS 7	19.9% 292	29.6% 131	35.0% 157	43.8% 154	22.0% 83
KSS 8	7.0% 102	19.7% 87	21.0% 94	27.3% 96	31.0% 117
KSS 9	1.9% 28	3.6% 16	9.4% 42	11.9% 42	36.5% 138

EEG (manual)
MAE: 1.12, Accuracy: 28.12%

	KSS 1-5	KSS 6	KSS 7	KSS 8	KSS 9
KSS 1-5	24.1% 354	14.0% 62	8.0% 36	3.4% 12	1.3% 5
KSS 6	40.0% 587	32.4% 143	26.3% 118	14.2% 50	14.6% 55
KSS 7	25.1% 368	33.7% 149	37.7% 169	38.6% 136	25.9% 98
KSS 8	9.2% 135	16.3% 72	21.4% 96	36.9% 130	39.2% 148
KSS 9	1.6% 23	3.6% 16	6.5% 29	6.8% 24	19.0% 72

EOG+EEG (manual)
MAE: 0.97, Accuracy: 33.72%

	KSS 1-5	KSS 6	KSS 7	KSS 8	KSS 9
KSS 1-5	40.4% 593	19.0% 84	12.1% 54	3.1% 11	5.0% 19
KSS 6	41.3% 606	36.9% 163	33.5% 150	21.9% 77	16.4% 62
KSS 7	13.3% 195	30.1% 133	33.9% 152	44.6% 157	26.2% 99
KSS 8	4.0% 58	10.4% 46	12.3% 55	23.3% 82	38.9% 147
KSS 9	1.0% 15	3.6% 16	8.3% 37	7.1% 25	13.5% 51

Figure 3: Summary of results from the regression analyses. The continuous regression output was rounded to the nearest class label when compiling the confusion matrices and when calculating the accuracy. The top row is for CNN-LSTM models with time-series input, and the bottom row for the shallow feedforward networks with manually extracted features based on domain knowledge.

The results from the binary classification models are shown as ROC curves in Figure 4 (left), and the performance at the optimal threshold is summarized in Table 4. The best overall performance was achieved with the CNN-LSTM model using both EEG and EOG as input. The sensitivity, i.e. the ability to correctly detect drivers who rated themselves as sleepy, was however better when only EEG data was used as input to the deep net.

When evaluating the regression models for binary classification, the CNN-LSTM model with only EOG data as input performed consistently better than the other models, as seen in Figure 4 (right) and in Table 5. Furthermore, as seen by comparing Table 4 and Table 5, this model even performed better than the best of the binary models, the CNN-LSTM model with EOG and EEG as input, for all performance metrics.

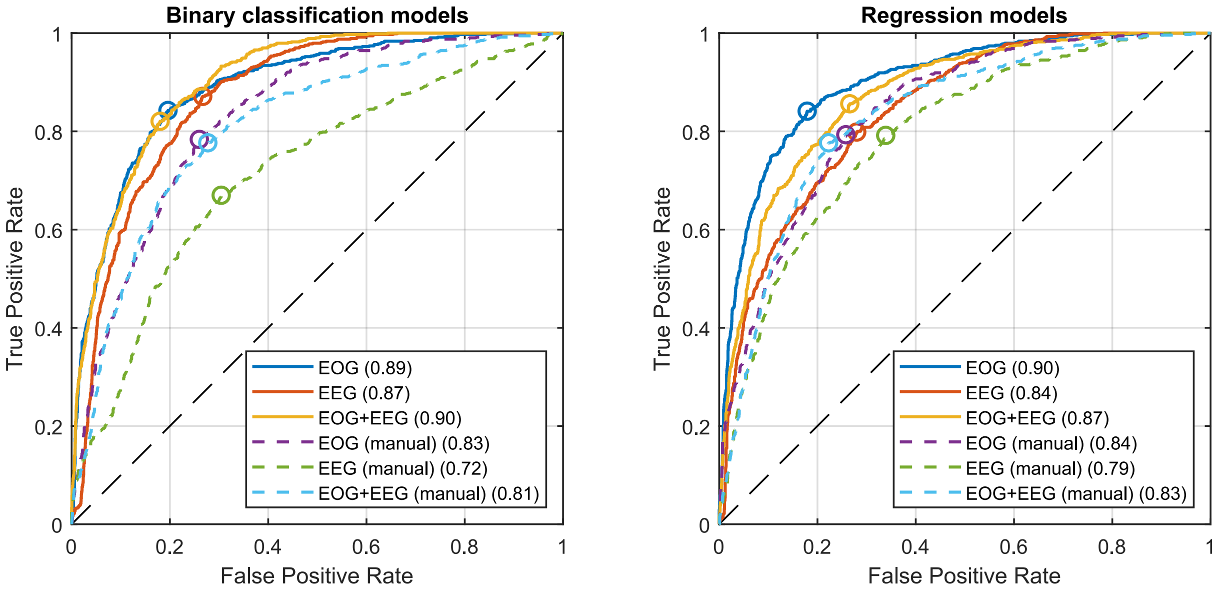


Figure 4: ROC curves for binary classification models (left) and regression models when used for binary classification (right). The circles mark the optimal points that maximize Youden's index for each model. The area under the ROC curves is given in the figure legend.

Table 4: Performance of the binary classification models. "Manual" refers to the shallow feedforward networks with manual feature extraction based on domain knowledge. The best model, per performance metric, is marked in bold.

	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1 (%)	AUC	Trainable parameters
EOG	81.4	84.2	80.4	69.9	0.89	157,762
EEG	76.9	**87.1**	73.3	65.8	0.87	158,082
EOG+EEG	**82.0**	82.1	**82.0**	**70.0**	**0.90**	158,242
EOG (manual)	75.2	78.4	74.0	63.6	0.83	3,074
EEG (manual)	68.9	67.0	69.6	54.3	0.72	6,146
EOG+EEG (manual)	73.8	77.7	72.3	62.1	0.81	9,026

Table 5: Performance of the regression models when the continuous output is discretized as alert or sleepy as in the binary classification case. "Manual" refers to the shallow feedforward networks with manual feature extraction based on domain knowledge. The best model, per performance metric, is marked in bold.

	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1 (%)	AUC	Trainable parameters
EOG	**82.6**	84.1	**82.1**	**71.2**	**0.90**	157,633
EEG	74.0	79.9	72.0	61.1	0.84	157,953
EOG+EEG	76.5	**85.6**	73.4	65.1	0.87	158,113
EOG (manual)	75.6	79.3	74.2	64.3	0.84	3,009
EEG (manual)	69.7	79.2	66.1	59.1	0.79	6,081
EOG+EEG (manual)	77.7	77.7	77.7	65.9	0.83	8,961

To investigate if the number of data segments were enough to train the deep network architectures, AUC was calculated as a function of an increasing number of training examples. The model used was the binary CNN-LSTM network with both EOG and EEG time series data as input. The data was shuffled before all the training sessions. Class weights were used to compensate for imbalances in the dataset. The results show that after an initial rapid increase in performance, the AUC continues to increase until the maximum number of available data segments is reached, see Figure 5. This indicates that adding more data would help improving the classifier further.

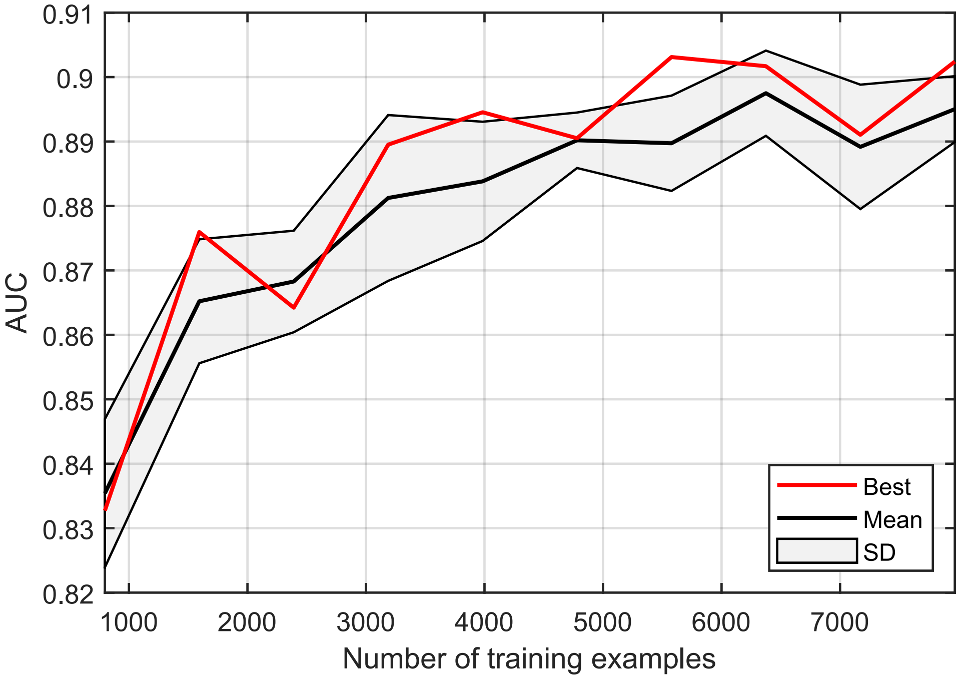


Figure 5: AUC for EOG+EEG CNN-LSTM model as a function of the amount of training data. 10 models were trained and the mean (black line) and standard deviation (grey shading) of AUC on test data is shown. The red line indicates the test AUC for the model with the highest AUC on validation data, i.e. the unbiased performance of the models.

4 Discussion

A driver sleepiness classification system based on a CNN-LSTM classifier, trained on data with high ecological validity, has been tested and evaluated. The MAE on the test dataset was 0.78. Binary classification accuracy on the test dataset was 82.6% for the regression model and 82.0% for the model that was trained specifically for the binary classification task. These results were better than what was achieved with a shallow neural network fed with manually extracted features (MAE 0.97, accuracy 77.7% in the regression case, 75.2% in the classification case). In general, data from the eyes (EOG) were more informative for driver sleepiness classification than data from the brain (EEG), and a combined input with both EEG and EOG did not always result in better performance.

When evaluating the regression models for binary classification (KSS ≤ 6 vs. KSS ≥ 8), the CNN-LSTM model with only EOG input substantially outperformed the other regression models. Even more surprising, the same model also outperformed the best of the binary models, i.e. the models that were trained specifically for the binary classification task. This indicates that the implicit order of the KSS labels hold important information that should not be discarded by aggregating them into an alert and a sleepy class. It also indicates that the information content is higher in the EOG signal than the EEG signal when considering the order of the KSS labels. This should be considered when designing future studies.

Including the feature extraction step in the machine learning model using deep learning provided higher classification performance compared to using manually defined domain knowledge-based features. This indicates that the deep features picked up information relevant for sleepiness scoring that was not picked up by the manually defined sleepiness indicators. To investigate the feature extraction capabilities of the CNN-LSTM network and to get a grasp of the extra information that was picked up by the deep features, outputs from the intermediate representations of the data were visualised. It was noticed that one intermediate layer represented eye blinks. Another early layer represented quick changes in amplitude between negative and positive peaks (i.e. saccades). Later layers seem to represent aggregates of several simple features, such as clusters of rapid blinking. These preliminary findings are well in line with the findings of Massoz (2019), who found inner representations of closed eyes, opened eyes, droopy eyes, long blinks, slow eyelid closures in a CNN trained to classify driver sleepiness based on video recordings.

The architecture of the CNN-LSTM network is based on previous work on sleep stage classification (Biswal et al., 2017, Stephansen et al., 2018, Biswal et al., 2018). The amount of possible combinations of layer types, activation functions, optimization function etc. is huge and all possible solutions have not been investigated. Future research should use the results from this work as a baseline and conduct a more systematic study of different model architectures. The procedure followed here was to start with a few layers and then gradually increase the complexity of the network. Using one convolutional layer alone in the feature extraction part of the network resulted in a classifier where most of the sleepy examples were classified as alert. As the number of layers were increased, the robustness of the classifier and the reproducibility of the results increased. Adding LSTM to the CNN network, as suggested by (Biswal et al., 2017), is reasonable from a theoretical point of view, as LSTM takes into consideration the time dependency in the signals. The added complexity of the classifier has greater potential to reach higher accuracy, but at the same time it is likely to require more training examples. We also investigated using separate CNN-networks for the EOG and EEG data types, as proposed by Stephansen et al. (2018), but no significant performance improvement was found compared to using EOG and EEG data in channels to the same CNN.

With 269 drivers and 1187 driving sessions, this is probably the largest labelled driver sleepiness dataset in the world with sleep deprived participants driving in an ecologically valid setting. Even so, this dataset is still limited. Figure 5 shows that classification performance increase as a function of the number of training examples and the graph indicates that performance will continue to increase if more training examples are provided. As a comparison, the clinical datasets that are used to train sleep stage classification algorithms typically consist of thousands of patients (Biswal et al., 2018, Stephansen et al., 2018). On the positive side, these results indicate that deep learning for driver sleepiness detection has not yet reached its full potential. Future research should exploit the possibility to pretrain especially the EEG subnetworks with data from medical sleep databases.

There is no truly objective ground truth of driver sleepiness that can be used when training supervised algorithms. There are alternatives to the subjective KSS ratings used here, including EEG measurements (Gillberg et al., 1996, Picot et al., 2012), blink durations (Schleicher et al., 2008), eye aperture (Kozak et al., 2005), reaction time tests (Massoz et al., 2018) and expert ratings based on observations (Awais et al., 2014, Rodriguez-Ibañez et al., 2012). However, reaction time tests are difficult to administer in real-road driving settings and video-based expert ratings have been found to be unreliable since sleepiness is often confused with boredom or underload (Ahlstrom et al., 2015). Physiology-based ground truths suffer from intra- as well and inter-individual differences and small effect sizes (Sparrow et al., 2018, Sandberg et al., 2011a, Åkerstedt et al., 2010, e.g. Caffier et al., 2003, Campagne et al., 2005, Ingre et al., 2006b, Schleicher et al., 2008, Papadelis et al., 2007). Regardless, both EEG and EOG were used as input data in the present paper and could thus not be used as the target label as well. The main drawbacks with KSS are that the subjective feeling does not always reflect the actual sleepiness level (Van Dongen et al., 2003), that repeated reporting can have an alerting effect (Kaida et al., 2007), and that participants may interpret the levels of KSS differently, leading to label noise. On the positive side, KSS correlates with lane departures and has been found to be the measure of driver sleepiness least affected by inter-individual variations (Åkerstedt et al., 2014). In addition, a system based on self-ratings will likely have high acceptance since it detects sleepiness in a way that match the drivers' expectations.

There are several limitations to the presented sleepiness detection system. *First*, using physiological data as input to the classifier limits its practical usefulness as obtrusive electrodes will never be accepted by the drivers. Instead, the developed classifier demonstrates the possibility to estimate sleepiness based on sensor data. Such a system can be used as a benchmark when developing future unobtrusive systems. It is positive that the performance when using only EOG as input was almost as good as when using EOG and EEG combined, and even better in some cases, because ocular parameters can be extracted from a video stream which would facilitate contact free driver sleepiness detection systems (Massoz et al., 2018, Schmidt et al., 2018). *Second*, the sleepiness dataset contain data recorded in naturalistic real-world settings as well as in an advanced moving-base simulator. It is well known that the progression from alert to sleepy is more rapid, and the absolute levels are generally higher, in a simulator environment (Fors et al., 2018, Hallvig et al., 2013). However, the physiological signs of sleepiness are similar. Given the need for large datasets when developing deep learning models, it was decided that the benefit of increasing the number of training examples outweighed the potential limitations of merging data from the two experimental settings. *Third*, the developed algorithm has not been benchmarked against previously presented sleepiness detection algorithms. In many cases, algorithms in the literature makes use of physiological signals that are not available in our database (muscle tension from the neck, respiration, 30-channel EEG etc.). In other cases, the developed algorithms are not subject-independent, or class imbalances have not been taken into consideration. One study, Mårtensson et al. (2019), used a subset of the present dataset and achieved a subject-independent accuracy of 84.0%, a sensitivity of 41.4% and a specificity of 93.1% based on a classic machine learning pipeline with feature selection and a random forest classifier. In comparison, the CNN-LSTM network developed here does not show this severe drop in performance for sensitivity. Further benchmarking, including comparisons with other network architectures, will be pursued in future research. *Fourth*, a system capable of preventing fatigue related crashes is not only dependent on a robust and accurate fatigue detection, but also on the effectiveness of the countermeasure that is then used to convince the driver to act in order to prevent an incident. The type of countermeasure that is needed depends on the type of fatigue that the driver is experiencing. Fatigue due to underload can be countered by doing something else for a while, fatigue due to overload can be remedied by a short break, while sleep-related fatigue can only be countered by actual sleep.

Future works should aim to further exploit the natural order of the KSS labels. This inherent structure was here exploited by imposing a regression structure on the output from the developed networks. This approach could be extended by also considering the transitions between different KSS levels, as well as the transition between alert and sleepy at KSS level 7. An approach could be to have overlapping training examples to handle these transitions between KSS levels. One could also elaborate on using manually extracted features and feed this data directly into a LSTM network to learn deeper eye movement and blink patterns.

5 Conclusions

The developed subject-independent CNN-LSTM classifier reached a binary classification accuracy of 82.0% and a mean absolute error of 0.78 in the regression case. Data from the eyes (EOG) were more informative for driver sleepiness detection than data from the brain (EEG). Combining EOG and EEG improved the performance for some models, but the gain was very limited. The deep CNN-LSTM network performed better than a shallow net fed with manually extracted sleepiness indicators. This shows that the deep network can extract information that is not captured by experts' domain knowledge.

6 Acknowledgements

This research is a collaborative effort that would not have been possible without all the hard work that went into the 12 experiments. We are especially indebted to A Anund and C Fors at VTI, T Åkerstedt and G Kecklund at the Stress Research Institute, Ingemar Fredriksson at Linköping University, and all of our project partners at Volvo Car Corporation, Autoliv Inc., former Saab Automobile AB, and Smart Eye AB.

This work was supported by the ADAS&ME project, which is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 688900.

7 References

AHLSTROM, C., FORS, C., ANUND, A. & HALLVIG, D. 2015. Video-based observer rated sleepiness versus self-reported subjective sleepiness in real road driving. *European Transport Research Review,* 7**,** 38.

ANUND, A., FORS, C., HALLVIG, D., AKERSTEDT, T. & KECKLUND, G. 2013. Observer rated sleepiness and real road driving: an explorative study. *PloS one,* 8**,** e64782.

ANUND, A., KECKLUND, G., VADEBY, A., HJÄLMDAHL, M. & ÅKERSTEDT, T. 2008. The alerting effect of hitting a rumble strip--a simulator study with sleepy drivers. *Accident analysis and prevention,* 40**,** 1970-1976.

APPARIES, R. J., RINIOLO, T. C. & PORGES, S. W. 1998. A psychophysiological investigation of the effects of driving longer-combination vehicles. *Ergonomics,* 41**,** 581-592.

AWAIS, M., BADRUDDIN, N. & DRIEBERG, M. 2014. A non-invasive approach to detect drowsiness in a monotonous driving environment. *Proceedings of the IEEE TENCON Region,* 10.

BALANDONG, R. P., AHMAD, R. F., SAAD, M. N. M. & MALIK, A. S. 2018. A Review on EEG-Based Automatic Sleepiness Detection Systems for Driver. *IEEE Access,* 6**,** 22908-22919.

BARUA, S., AHMED, M. U., AHLSTRÖM, C. & BEGUM, S. 2019. Automatic driver sleepiness detection using EEG, EOG and contextual information. *Expert systems with applications,* 115**,** 121-135.

BEKIARIS, E., NIKOLAOU, S., PETERS, B. & ANUND, A. 2005. Driver fatigue monitoring, detection & warning: AWAKE project final results. *ITS Europe***,** 1-3.

BIOULAC, S., FRANCHI, J.-A. M., ARNAUD, M., SAGASPE, P., MOORE, N., SALVO, F. & PHILIP, P. 2017. Risk of motor vehicle accidents related to sleepiness at the wheel: a systematic review and meta-analysis. *Sleep,* 40.

BISWAL, S., KULAS, J., SUN, H., GOPARAJU, B., WESTOVER, M. B., BIANCHI, M. T. & SUN, J. 2017. SLEEPNET: automated sleep staging system via deep learning. *arXiv preprint arXiv:1707.08262*.

BISWAL, S., SUN, H., GOPARAJU, B., WESTOVER, M. B., SUN, J. & BIANCHI, M. T. 2018. Expert-level sleep scoring with deep neural networks. *Journal of the American Medical Informatics Association,* 25**,** 1643-1650.

CAFFIER, P. P., ERDMANN, U. & ULLSPERGER, P. 2003. Experimental evaluation of eye-blink parameters as a drowsiness measure. *Eur J Appl Physiol,* 89**,** 319-25.

CAMPAGNE, A., PEBAYLE, T. & MUZET, A. 2005. Oculomotor changes due to road events during prolonged monotonous simulated driving. *Biol Psychol,* 68**,** 353-68.

CHAI, R., LING, S. H., SAN, P. P., NAIK, G. R., NGUYEN, T. N., TRAN, Y., CRAIG, A. & NGUYEN, H. T. 2017. Improving eeg-based driver fatigue classification using sparse-deep belief networks. *Frontiers in neuroscience,* 11**,** 103.

CHEE, M. W. 2015. Limitations on visual information processing in the sleep-deprived brain and their underlying mechanisms. *Current opinion in behavioral sciences,* 1**,** 56-63.

CHOWDHURY, A., SHANKARAN, R., KAVAKLI, M. & HAQUE, M. M. 2018. Sensor Applications and Physiological Features in Drivers' Drowsiness Detection: A Review. *IEEE Sensors Journal,* 18**,** 3055-3067.

CONNOR, J., NORTON, R., AMERATUNGA, S., ROBINSON, E., CIVIL, I., DUNN, R., BAILEY, J. & JACKSON, R. 2002. Driver sleepiness and risk of serious injury to car occupants: population based case control study. *BMJ,* 324**,** 1125-1128A.

DE NAUROIS, C. J., BOURDIN, C., BOUGARD, C. & VERCHER, J. L. 2018. Adapting artificial neural networks to a specific driver enhances detection and prediction of drowsiness. *Accident Analysis & Prevention,* 121**,** 118-128.

FERNÁNDEZ, A., USAMENTIAGA, R., CARÚS, J. & CASADO, R. 2016. Driver Distraction Using Visual-Based Sensors and Algorithms. *Sensors,* 16**,** 1805.

FORS, C., AHLSTRÖM, C. & ANUND, A. 2018. A comparison of driver sleepiness in the simulator and on the real road. *Journal of Transporrtation, Safety and Security,* 10**,** 72-87.

FU, R., WANG, H. & ZHAO, W. 2016. Dynamic driver fatigue detection using hidden Markov model in real driving condition. *Expert Systems with Applications,* 63**,** 397-411.

GILLBERG, M., KECKLUND, G. & AKERSTEDT, T. 1996. Sleepiness and performance of professional drivers in a truck simulator--comparisons between day and night driving. *J Sleep Res,* 5**,** 12-5.

GOLZ, M., SOMMER, D., CHEN, M., MANDIC, D. & TRUTSCHEL, U. 2007. Feature fusion for the detection of microsleep events. *Journal of VLSI Signal Processing Systems for Signal Image and Video Technology,* 49**,** 329-342.

GOLZ, M., SOMMER, D. & KRAJEWSKI, J. 2016. Prediction of immediately occurring microsleep events from brain electric signals. *Current Directions in Biomedical Engineering,* 2**,** 149-153.

GOLZ, M., SOMMER, D., TRUTSCHEL, U., SIROIS, B. & EDWARD, D. 2010. Evaluation of fatigue monitoring technologies. *Somnologie,* 14**,** 187-199.

HALLVIG, D., ANUND, A., FORS, C., KECKLUND, G. & AKERSTEDT, T. 2014. Real driving at night - Predicting lane departures from physiological and subjective sleepiness. *Biological Psychology,* 101**,** 18-23.

HALLVIG, D., ANUND, A., FORS, C., KECKLUND, G., KARLSSON, J. G., WAHDE, M. & ÅKERSTEDT, T. 2013. Sleepy driving on the real road and in the simulator--A comparison. *Accident analysis and prevention,* 50**,** 44-50.

HU, S., ZHENG, G. & PETERS, B. 2013. Driver fatigue detection from electroencephalogram spectrum after electrooculography artefact removal. *IET Intelligent Transport Systems,* 7**,** 105-113.

INGRE, M., PETERS, B., ANUND, A., KECKLUND, G. & PICKLES, A. 2006a. Subjective sleepiness and accident risk avoiding the ecological fallacy. *Journal of sleep research,* 15**,** 142-148.

INGRE, M., ÅKERSTEDT, T., PETERS, B., ANUND, A. & KECKLUND, G. 2006b. Subjective sleepiness, simulated driving performance and blink duration: examining individual differences. *Journal of Sleep Research,* 15**,** 47-53.

JAMMES, B., SHARABTY, H. & ESTEVE, D. 2008. Automatic EOG analysis: A first step toward automatic drowsiness scoring during wake-sleep transitions *Somnologie - Schlafforschung und Schlafmedizin,* 12**,** 227-232.

KAIDA, K., ÅKERSTEDT, T., KECKLUND, G., NILSSON, J. P. & AXELSSON, J. 2007. The effects of asking for verbal ratings of sleepiness on sleepiness and its masking effects on performance. *Clinical Neurophysiology,* 118**,** 1324-1331.

KECKLUND, G. & ÅKERSTEDT, T. 1993. Sleepiness in long distance truck driving: An ambulatory EEG study of night driving. *Ergonomics,* 36**,** 1007-1017.

KHUSHABA, R. N., KODAGODA, S., LAL, S. & DISSANAYAKE, G. 2011. Driver drowsiness classification using fuzzy wavelet-packet-based feature-extraction algorithm. *IEEE Transactions on Biomedical Engineering,* 58**,** 121-131.

KONG, W., ZHOU, Z., JIANG, B., BABILONI, F. & BORGHINI, G. 2017. Assessment of driving fatigue based on intra/inter-region phase synchronization. *Neurocomputing,* 219**,** 474-482.

KOZAK, K., CURRY, R., GREENBERG, J., ARTZ, B., BLOMMER, M. & CATHEY, L. 2005. Leading Indicators of Drowsiness in Simulated Driving. *Human Factors and Ergonomics Society Annual Meeting Proceedings,* 49**,** 1917-1917.

KRAUSE, A. J., SIMON, E. B., MANDER, B. A., GREER, S. M., SALETIN, J. M., GOLDSTEIN-PIEKARSKI, A. N. & WALKER, M. P. 2017. The sleep-deprived human brain. *Nature Reviews Neuroscience,* 18**,** 404.

LEANDERSSON OLSSON, S. 2012. Drowsy Driver Project Final Report. Vinnova.

LI, G. & CHUNG, W. Y. 2015. A context-aware EEG headset system for early detection of driver drowsiness. *Sensors (Switzerland),* 15**,** 20873-20893.

LI, G., LEE, B. L. & CHUNG, W. Y. 2015. Smartwatch-Based Wearable EEG System for Driver Drowsiness Detection. *IEEE Sensors Journal,* 15**,** 7169-7180.

LIN, F. C., KO, L. W., CHUANG, C. H., SU, T. P. & LIN, C. T. 2012. Generalized EEG-based drowsiness prediction system by using a self-organizing neural fuzzy system. *IEEE Transactions on Circuits and Systems I: Regular Papers,* 59**,** 2044-2055.

LIU, C. C., HOSKING, S. G. & LENNÉ, M. G. 2009. Predicting driver drowsiness using vehicle measures: recent insights and future challenges. *Journal of safety research,* 40**,** 239-245.

LIU, J., ZHANG, C. & ZHENG, C. 2010. EEG-based estimation of mental fatigue by using KPCA–HMM and complexity parameters. *Biomedical Signal Processing and Control,* 5**,** 124-130.

MA, Y., CHEN, B., LI, R., WANG, C., WANG, J., SHE, Q., LUO, Z. & ZHANG, Y. 2019. Driving fatigue detection from EEG using a modified PCANet method. *Computational intelligence and neuroscience,* 2019.

MASSOZ, Q. 2019. *Non-invasive, automatic, and real-time characterization of drowsiness based on eye closure dynamics.* Université de Liège, Liège, Belgique.

MASSOZ, Q., VERLY, J. & VAN DROOGENBROECK, M. 2018. Multi-Timescale Drowsiness Characterization Based on a Video of a Driver's Face. *Sensors,* 18.

MU, Z., HU, J. & MIN, J. 2017. Driver fatigue detection system using electroencephalography signals based on combined entropy features. *Applied Sciences (Switzerland),* 7.

MÅRTENSSON, H., KEELAN, O. & AHLSTRÖM, C. 2019. Driver Sleepiness Classification Based on Physiological Data and Driving Performance From Real Road Driving. *IEEE Transactions on Intelligent Transportation Systems,* 20**,** 421-430.

PAPADELIS, C., CHEN, Z., KOURTIDOU-PAPADELI, C., BAMIDIS, P. D., CHOUVARDA, I., BEKIARIS, E. & MAGLAVERAS, N. 2007. Monitoring sleepiness with on-board electrophysiological recordings for preventing sleep-deprived traffic accidents. *Clinical Neurophysiology,* 118**,** 1906-1922.

PATEL, M., LAL, S., KAVANAGH, D. & ROSSITER, P. 2011. Applying neural network analysis on heart rate variability data to assess driver fatigue. *Expert systems with Applications,* 38**,** 7235-7242.

PERSSON, A., JONASSON, H., FREDRIKSSON, I., WIKLUND, U. & AHLSTRÖM, C. Heart Rate Variability for Driver Sleepiness Classification in Real Road Driving Conditions. 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2019. IEEE, 6537-6540.

PHILLIPS, R. O. & SAGBERG, F. 2013. Road accidents caused by sleepy drivers: Update of a Norwegian survey. *Accident Analysis & Prevention,* 50**,** 138-146.

PICOT, A., CHARBONNIER, S. & CAPLIER, A. 2012. On-line detection of drowsiness using brain and visual information. *IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans,* 42**,** 764-775.

RADUN, I., OHISALO, J., RADUN, J., WAHDE, M. & KECKLUND, G. 2013. Driver fatigue and the law from the perspective of police officers and prosecutors. *Transportation research part F: traffic psychology and behaviour,* 18**,** 159-167.

RADUN, I., WAHDE, M., INGRE, M., BENDERIUS, O. & KECKLUND, G. 2014. Driving while fatigued in slippery road conditions - a neglected issue. *22nd Congress of the European Sleep Research Society, 16-20 September, 2014, Tallin, Estonia.*

RAMZAN, M., KHAN, H. U., AWAN, S. M., ISMAIL, A., ILYAS, M. & MAHMOOD, A. 2019. A Survey on State-of-the-Art Drowsiness Detection Techniques. *IEEE Access,* 7**,** 61904-61919.

RODRIGUEZ-IBAÑEZ, N., GARCÍA-GONZALEZ, M. A., DE LA CRUZ, M. A. F., FERNÁNDEZ-CHIMENO, M. & RAMOS-CASTRO, J. 2012. Changes in heart rate variability indexes due to drowsiness in professional drivers measured in a real environment. *Computing in Cardiology,* 39**,** 913-916.

ROY, Y., BANVILLE, H., ALBUQUERQUE, I., GRAMFORT, A., FALK, T. H. & FAUBERT, J. 2019. Deep learning-based electroencephalography analysis: a systematic review. *Journal of neural engineering,* 16**,** 051001.

SAHAYADHAS, A., SUNDARAJ, K. & MURUGAPPAN, M. 2012. Detecting driver drowsiness based on sensors: A review. *Sensors (Switzerland),* 12**,** 16937-16953.

SANDBERG, D., ANUND, A., FORS, C., KECKLUND, G., KARLSSON, J. G., WAHDE, M. & ÅKERSTEDT, T. 2011a. The Characteristics of Sleepiness During Real Driving at Night-A Study of Driving Performance, Physiology and Subjective Experience. *SLEEP,* 34**,** 1317-1325.

SANDBERG, D., ÅKERSTEDT, T., ANUND, A., KECKLUND, G. & WAHDE, M. 2011b. Detecting Driver Sleepiness Using Optimized Nonlinear Combinations of Sleepiness Indicators. *IEEE Transactions on Intelligent Transportation Systems,* 12**,** 97-108.

SCHLEICHER, R., GALLEY, N., BRIEST, S. & GALLEY, L. 2008. Blinks and saccades as indicators of fatigue in sleepiness warnings: looking tired? *Ergonomics,* 51**,** 982-1010.

SCHMIDT, J., LAAROUSI, R., STOLZMANN, W. & KARRER-GAUß, K. 2018. Eye blink detection for different driver states in conditionally automated driving and manual driving using EOG and a driver camera. *Behavior research methods,* 50**,** 1088-1101.

SCHWARZ, J. F., INGRE, M., FORS, C., ANUND, A., KECKLUND, G., TAILLARD, J., PHILIP, P. & ÅKERSTEDT, T. 2012. In-car countermeasures open window and music revisited on the real road: popular but hardly effective against driver sleepiness. *Journal of Sleep Research,* 21**,** 595-599.

SHAHID, A., SHEN, J. & SHAPIRO, C. M. 2010. Measurements of sleepiness and fatigue. *Journal of psychosomatic research,* 69**,** 81-89.

SIKANDER, G. & ANWAR, S. 2018. Driver fatigue detection systems: A review. *IEEE Transactions on Intelligent Transportation Systems*.

SILVEIRA, C. S., CARDOSO, J. S., LOURENÇO, A. L. & AHLSTRÖM, C. 2019. Importance of subject-dependent classification and imbalanced distributions in driver sleepiness detection in realistic conditions. *IET Intelligent Transport Systems,* 13**,** 347-355.

SIMON, M., SCHMIDT, E. A., KINCSES, W. E., FRITZSCHE, M., BRUNS, A., AUFMUTH, C., BOGDAN, M., ROSENSTIEL, W. & SCHRAUF, M. 2011. EEG alpha spindle measures as indicators of driver fatigue under real traffic conditions. *Clinical Neurophysiology,* 122**,** 1168-1178.

SONG, Y., WANG, D., YUE, K. & ZHENG, N. 68-3: DeepFatigueNet: A Model for Automatic Visual Fatigue Assessment Based on Raw Single-Channel EEG. SID Symposium Digest of Technical Papers, 2019. Wiley Online Library, 965-968.

SPARROW, A. R., LAJAMBE, C. M. & VAN DONGEN, H. P. A. 2018. Drowsiness measures for commercial motor vehicle operations. *Accident Analysis & Prevention,* 126**,** 146-159.

STEPHANSEN, J. B., OLESEN, A. N., OLSEN, M., AMBATI, A., LEARY, E. B., MOORE, H. E., CARRILLO, O., LIN, L., HAN, F. & YAN, H. 2018. Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. *Nature communications,* 9**,** 1-15.

TRAN, Y., WIJESURIYA, N., TARVAINEN, M., KARJALAINEN, P. & CRAIG, A. 2009. The Relationship Between Spectral Changes in Heart Rate Variability and Fatigue. *JOURNAL OF PSYCHOPHYSIOLOGY,* 23**,** 143-151.

VAN DEN BERG, J., NEELY, G., WIKLUND, U., LANDSTRÖM, U., STRÅLNINGSVETENSKAPER, MEDICINSK, F., FOLKHÄLSA OCH KLINISK, M. & UMEÅ, U. 2005. Heart rate variability during sedentary work and sleep in normal and sleep-deprived states. *Clinical Physiology and Functional Imaging,* 25**,** 51-51.

VAN DONGEN, H. P. A., MAISLIN, G., MULLINGTON, J. M. & DINGES, D. F. 2003. The cumulative cost of additional wakefulness: dose-response effects on neurobehavioral functions and sleep physiology from chronic sleep restriction and total sleep deprivation. *Sleep,* 26**,** 117-126.

VICENTE, J., LAGUNA, P., BARTRA, A. & BAILÓN, R. 2016. Drowsiness detection using heart rate variability. *Medical & Biological Engineering & Computing,* 54**,** 927-937.

WORLD HEALTH ORGANIZATION 2018. Global status report on road safety 2018.

YANG, G., LIN, Y. & BHATTACHARYA, P. 2010. A driver fatigue recognition model based on information fusion and dynamic Bayesian network. *Information Sciences,* 180**,** 1942-1954.

YOUDEN, W. J. 1950. Index for rating diagnostic tests. *Cancer,* 3**,** 32-35.

ZENG, H., YANG, C., DAI, G., QIN, F., ZHANG, J. & KONG, W. 2018. EEG classification of driver mental states by deep learning. *Cognitive neurodynamics,* 12**,** 597-606.

ÅKERSTEDT, T. 2000. Consensus statement: fatigue and accidents in transport operations. *Journal of sleep research,* 9**,** 395-395.

ÅKERSTEDT, T., ANUND, A., AXELSSON, J. & KECKLUND, G. 2014. Subjective sleepiness is a sensitive indicator of insufficient sleep and impaired waking function. *Journal of Sleep Research,* 23**,** 240-252.

ÅKERSTEDT, T. & GILLBERG, M. 1990. Subjective and objective sleepiness in the active individual. *International Journal of Neuroscience,* 52**,** 29-37.

ÅKERSTEDT, T., INGRE, M., KECKLUND, G., ANUND, A., SANDBERG, D., WAHDE, M., PHILIP, P. & KRONBERG, P. 2010. Reaction of sleepiness indicators to partial sleep deprivation, time of day and time on task in a driving simulator--the DROWSI project. *Journal of sleep research,* 19**,** 298-309.

ÅKERSTEDT, T., PETERS, B., ANUND, A. & KECKLUND, G. 2005. Impaired alertness and performance driving home from the night shift: a driving simulator study. *Journal of sleep research,* 14**,** 17-20.