Expert Systems With Applications 115 (2019) 121–135

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Automatic driver sleepiness detection using EEG, EOG and contextual information

Shaibal Barua[a,*], Mobyen Uddin Ahmed[a,*], Christer Ahlström[b,c], Shahina Begum[a]

[a] School of Innovation Design and Engineering, Mälardalen University, Västerås, SE 72123, Sweden
[b] The Swedish National Road and Transport Research Institute (VTI), Linköping, SE 58195, Sweden
[c] Department of Biomedical Engineering, Linköping University, Linköping, SE 58183, Sweden

ARTICLE INFO

Article history:
Received 5 March 2018
Revised 26 July 2018
Accepted 27 July 2018
Available online 3 August 2018

Keywords:
Driver sleepiness
Electroencephalography
Electrooculography
Contextual information
Machine learning

ABSTRACT

The many vehicle crashes that are caused by driver sleepiness each year advocates the development of automated driver sleepiness detection (ADSD) systems. This study proposes an automatic sleepiness classification scheme designed using data from 30 drivers who repeatedly drove in a high-fidelity driving simulator, both in alert and in sleep deprived conditions. Driver sleepiness classification was performed using four separate classifiers: k-nearest neighbours, support vector machines, case-based reasoning, and random forest, where physiological signals and contextual information were used as sleepiness indicators. The subjective Karolinska sleepiness scale (KSS) was used as target value. An extensive evaluation on multiclass and binary classifications was carried out using 10-fold cross-validation and leave-one-out validation. With 10-fold cross-validation, the support vector machine showed better performance than the other classifiers (79% accuracy for multiclass and 93% accuracy for binary classification). The effect of individual differences was also investigated, showing a 10% increase in accuracy when data from the individual being evaluated was included in the training dataset. Overall, the support vector machine was found to be the most stable classifier. The effect of adding contextual information to the physiological features improved the classification accuracy by 4% in multiclass classification and by and 5% in binary classification.

1. Introduction

Driver sleepiness leads to worsened decision making, slower reaction times, and reduced attention to the forward roadway (Anderson & Horne, 2013; Dinges, 1995; NCSDR/NHTSA, 1998). As such, driver sleepiness is an imperative contributing factor to vehicle crashes and traffic injuries (Horne & Reyner, 1999; Philip et al., 2005). The National Highway Traffic Safety Administration (NHTSA) reports that 2.6% of crash fatalities in the USA in 2014 were due to driver sleepiness, and that 846 people died in those crashes (Lyles, 2015). The International Transport Forum at the OECD estimates that 20–30% of fatalities are due to driver sleepiness and fatigue (ITF, 2017). In Europe, a survey across nineteen countries have shown that the average prevalence of falling asleep while driving in the previous 2 years is 17%, and amongst those who fall asleep, the prevalence of sleep-related crashes is 7% (Gonçalves et al., 2015). The high prevalence and crash rates advocate the development of automated driver sleepiness detection (ADSD), which is the topic of this research work.

Much research has been conducted to determine indicators and measures of driver sleepiness (Borghini, Astolfi, Vecchiato, Mattia, & Babiloni, 2014; Curcio, Casagrande, & Bertini, 2001; Sparrow, LaJambe, & Van Dongen, 2018). These indicators are often grouped as vehicle-based, behavioural-based, physiological, or a combination thereof. Vehicle-based indictors account for the driver's ability to operate the vehicle in a reasonable and expected manner, in terms of variability in lateral position, appropriate distance to surrounding vehicles, etc. (Liu, Hosking, & Lenné, 2009; Sandberg, Åkerstedt, Anund, Kecklund, & Wahde, 2011). Behavioural measures incorporate information such as visual scanning and yawning (Anund, Fors, Hallvig, Åkerstedt, & Kecklund, 2013; Watling, Armstrong, & Radun, 2015). Physiological measures include sleepiness indicators based on electroencephalography (EEG), electrooculography (EOG), respiration, heart rate and heart rate variability (HRV) (Čolić, Marques, & Furht, 2014; Sahayadhas, Sundaraj, & Murugappan, 2012). This paper focuses on physiological indicators of sleepiness, but also incorporate contextual information such as light conditions, driving environment and time awake.

* Corresponding authors.
E-mail addresses: shaibal.barua@mdh.se (S. Barua), mobyen.ahmed@mdh.se (M.U. Ahmed), christer.ahlstrom@vti.se (C. Ahlström), shahina.begum@mdh.se (S. Begum).

https://doi.org/10.1016/j.eswa.2018.07.054
0957-4174/

EEG is the most commonly used physiological signal in automatic classification of sleep stages (Boostani, Karimzadeh, & Nami, 2017; Hassan & Bhuiyan, 2016a, 2016b, 2017), and it has also been used to estimate different sleepiness levels (Abeyratne, Vinayak, Hukins, & Duce, 2009; Akin, Kurt, Sezgin, & Bayram, 2008; Balandong, Ahmad, Saad, & Malik, 2018; Borghini et al., 2014; Jiao & Lu, 2017; Kar, Bhagat, & Routray, 2010). The frequency power of the EEG signal is typically measured since increased theta (θ) power (4–7 Hz) (Aeschbach et al., 1997; Cajochen, Brunner, Kräuchi, Graw, & Wirz-Justice, 1995) as well as increased alpha (α) power (8–12 Hz) indicate sleepiness, whereas increased beta (β) content (12–30 Hz) is a sign of alertness (Craig, Tran, Wijesuriya, & Nguyen, 2012). In a driving context, especially increased alpha band power has been found to be associated with sleepiness (Kecklund & Åkerstedt, 1993; Simon et al., 2011). Longer blink durations (Åkerstedt, Peters, Anund, & Kecklund, 2005; Häkkänen, LicPsych, Summala, Partinen, Tiihonen, & Silvo, 1999; Schleicher, Galley, Briest, & Galley, 2008) and slow eye movements (Kurt, Sezgin, Akin, Kirbas, & Bayram, 2009), measured via EOG, are other indictors of sleepiness. Also HRV contributes with information related to sleepiness (Vicente, Laguna, Bartra, & Bailón, 2016), but it is more commonly used to measure driver fatigue and stress (s Begum, Ahmed, Funk, & Filla, 2012; Begum, Barua, Filla, & Ahmed, 2014).

Contextual information tries to account for some of the many confounding factors that affect sleep and sleepiness (Åkerstedt et al., 2004; Cooke & Ancoli-Israel, 2011; Menefee et al., 2000; Reyner, Wells, Mortlock, & Horne, 2012). In a driving setting, environmental factors such as light conditions, road curvature and traffic density affect sleepiness and fatigue (Ahlström, Anund, Fors, & Åkerstedt, 2018b; Figueiro, Bullough, & Rea, 2007; Rüger, Gordijn, Beersma, de Vries, & Daan, 2006). Another category of contextual factors that affect sleepiness are the time of the day, the time awake and prior sleep (Åkerstedt, Connor, Gray, & Kecklund, 2008). Physiological indicators vary from person to person (Sparrow et al., 2018) so by adding contextual information to the ADSD system, the idea is that the effect of some of these intra-individual variations can be countered.

A common approach in ADSD systems is to use data driven approaches based on several data sources (Chacon-Murguia & Prieto-Resendiz, 2015; Golz, Sommer, Trutschel, Sirois, & Edwards, 2010; Jacobé de Naurois, Bourdin, Stratulat, Diaz, & Vercher, 2017; Li & Chung, 2018; Mårtensson, Keelan, & Ahlström, 2018; Sahayadhas et al., 2012). Machine learning methods that have been used for ADSD includes support vector machines (SVM) (Chen, Zhao, Ye, Zhang, & Zou, 2017; Chui, Tsang, Chi, Wu, & Ling, 2015; Hu & Zheng, 2009; Yeo, Li, Shen, & Wilder-Smith, 2009), linear discriminant analysis (Vicente et al., 2016), artificial neural networks (Dwivedi, Biswaranjan, & Sethi, 2014; Garcés Correa, Orosco, & Laciar, 2014; Ma, Murphey, & Zhao, 2015), logistic regression (Babaeian, Bhardwaj, Esquivel, & Mozumdar, 2016), and k-means clustering (Gurudath & Riley, 2014). However, most of these ADSD systems have been developed using rather small datasets that does not allow for proper validation of the sophistic methods that are used. Also, many studies are not actually using data from sleepy drivers. Instead "sleepy" data is invoked by driving for about an hour in a monotonous setting. Such an experimental design gives rise to fatigue caused by under-stimulation rather than to physiological sleepiness.

The aim of this paper is to develop an ADSD system that exploits physiological as well as contextual information. A subjective measure of sleepiness, the *Karolinska Sleepiness Scale* (KSS) (Åkerstedt & Gillberg, 1990), is used as a target value in a supervised machine learning setup. Four different machine learning algorithms, k-nearest neighbour (KNN), support vector machine (SVM), case-based reasoning (CBR), and Random Forest (RF), are investigated. In addition, three feature selection algorithms are tested and evaluated for obtaining optimal feature subsets. An extensive evaluation has been performed to compare the effectiveness of these machine learning algorithms for both multiclass and binary classification.

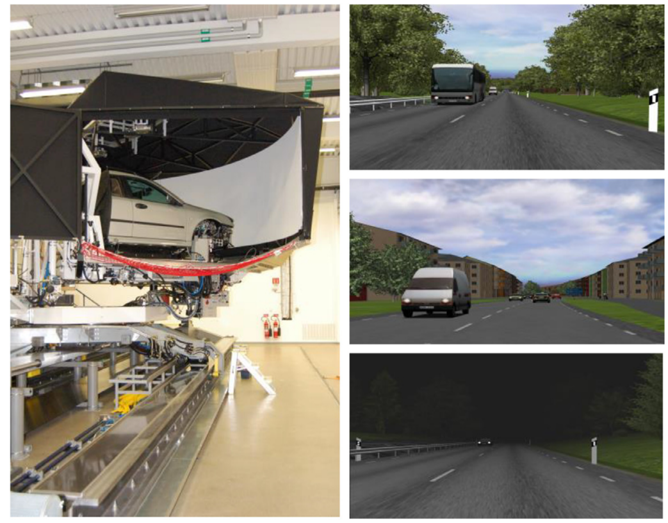

Fig. 1. The VTI Driving Simulator III (left) and examples of the simulated rural road in daylight, the suburban road, and the rural road in darkness (right). The right-hand side of the projector dome has been removed to show the car body and the projection screen.

2. Data collection and study procedure

The study took place in a high-fidelity moving-base driving simulator (VTI Driving Simulator III[1]) at the Swedish National Road and Transport Research Institute (VTI), see Fig. 1. By moving, rotating, or tilting the part of the simulator containing the car body and projector screens, acceleration and deceleration forces in either direction can be simulated. A vibration table enables a simulation of road surface contact. Six projectors are used for visualization of the frontal view with a horizontal field of view of 120°. Three LCD-displays are used as rear-view mirrors. The simulator has been validated in terms of driver sleepiness (Fors, Ahlstrom, & Anund, 2018; Hallvig et al., 2013).

Thirty participants were randomly selected from the Swedish register of vehicle owners. Inclusion criteria, in order to avoid confounding with known factors sensitive to sleepiness, were: young age (18–25 years old), male gender, body mass index less than 30, no shift workers, self-reported evening type, no sleep disorder, no extremes in terms of self-reported personalities (extrovert or introvert), and self-reported normal sensitivity to stressful situations. To avoid simulator sickness, only participants who were not prone to motion sickness were included. Before arrival, the participants were requested to avoid alcohol for 72 h and to abstain from nicotine and caffeine for 3 h before driving. The participants signed an informed consent form, and the study was approved by the regional ethics committee at Linköping University, Sweden (Dnr 2014/309-31).

The participants visited the laboratory on six separate occasions, three times during daytime and three times during nighttime. The day sessions (supposedly alert condition) were run between 12.30 h and 21.15 h and the night sessions (sleep deprived condition) were run between 22.00 h and 06.15 h. The participants were instructed to sleep for at least 7 h during the three days

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

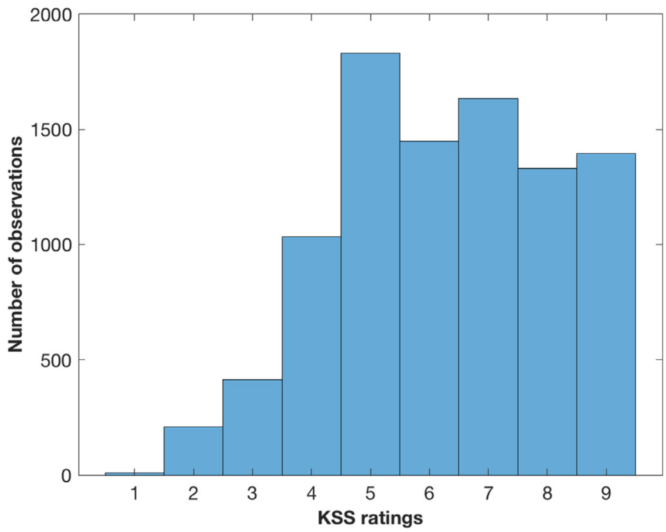

Fig. 2. Distribution of KSS ratings in the database.

before the trials, to go to bed no later than 24.00 h, and to get up no later than 09.00 h.

The participants drove 3 simulated scenarios on each of their six visits: (i) a rural road with a speed limit of 80 km/h in daylight, (ii) the same rural road in darkness, and (iii) a suburban road in daylight and a speed limit of 60 km/h, see Fig. 1. The order of the three scenarios was randomized between participants, but the order of the scenarios was held constant within participants to facilitate studies on intra-individual differences (these results will be reported elsewhere). Each scenario lasted for 30 min, with 1.5 h of rest in between each drive. In total, 540 drives were conducted (30 drivers x 6 occasions x 3 scenarios).

Physiological data were acquired using a multi-channel amplifier with active electrodes (g.HIamp, g.tec Medical Engineering GmbH, Austria). The electroencephalography (EEG) electrodes were positioned based on the 10–20 system providing a 30-channel recording. The EEG signals were band-pass filtered between 0.5 and 60 Hz using an 8th order Butterworth filter, and frequencies between 48 and 52 Hz were removed using a 4th order Butterworth notch filter. In addition, electrooculography (EOG) (horizontal with electrodes at the outer canthi and vertical with electrodes above/below the left eye) were also acquired. The drivers used KSS to rate their sleepiness status every fifth minute throughout the drives. KSS has nine anchored levels: 1 – extremely alert, 3 – alert, 5 – neither alert nor sleepy, 7 – sleepy, no effort to stay awake, and 9 – very sleepy, great effort to keep awake, fighting sleep. A text was shown on the projection screen every fifth minute asking the driver to rate his/her sleepiness level, and the driver responded by pressing a digit on a touch screen in the centre stack. The reported value corresponds to the average feeling during the past 5 min. The distribution of KSS ratings in the database is presented in Fig. 2.

3. Sleepiness classification scheme for experiments

The overall classification scheme for the experimental work is shown in Fig. 3. The scheme consists of five sub-modules: signal pre-processing, feature extraction, feature selection, sub-group creation, and finally sleepiness classification. The classification scheme is derived using features extracted from 1 min recordings (see 3.2). Four separate classifiers, KNN, SVM, CBR, and RF, were developed and trained using a) 10-fold cross-validation and b) leave-one-out (LOO) validation. Both these validation approaches were used since k-fold (k = 10) cross-validation is favourable when estimating the expected prediction error, whereas leave-one-out validation is almost unbiased (Elisseeff & Pontil, 2003). For 10-fold cross-validation, the dataset was randomly split into training and testing datasets, with 70% of the observations in the training set and 30% of the data in the testing set, which can be seen in Fig. 4. Note that observations from one individual driver can be included in both the training set and in the test data set, which will result in a subject-dependent classifier. For LOO validation, all observations from one driver were used as test data and the remaining observations were used for training. This approach will result in a generalized subject-independent classifier. To further investigate the influence of individual differences and how they impact the resulting classifiers, an additional LOO evaluation was conducted. Here, instead of leaving out observations from one participant as before, we left out observations from one (5 min) KSS rating.

10-fold cross-validation was used when training the models in both the LOO and 10-fold cross-validation evaluations.

3.1. Pre-processing

During driving, the driver dynamically engages in a series of actions such as turning the steering wheel, changing gears, scanning the forward road scene, looking at side and rear-view mirrors and looking over the shoulder. In addition, a sleepy driver tends to yawn, change body position and lean the head against the head rest (Anund et al., 2013). EEG data recorded while driving is therefore heavily contaminated with ocular, muscle and motion artifacts. Hence, before feature extraction, the acquired EEG signals were artifacts handled using the ARTE tool (Barua, Ahmed, Ahlstrom, Begum, & Funk, 2017). ARTE is an in-house developed tool which has been evaluated qualitatively by an expert in neurophysiology as well as by several quantitative measures.

3.2. Feature extraction

The time resolution of the target values, i.e. the KSS ratings, was one rating every fifth minute (this is a standard procedure when using KSS in a driving setting). However, for feature extraction, these five-minute segments were split into five one-minute segments and features were extracted from each of these one-minute segments. Extracting features from 1 min intervals not only provides a larger dataset compared to the 5 min level, but also reduces the sensitivity to small variations and corrective driving manoeuvres (Sandberg et al., 2011).

Data from 4 of the 30 participants had to be excluded due to poor data quality and inconsistency in the recordings. Hence, data from 312 drives (26 drivers x 5 occasions x 3 scenarios) were used in this study, excluding the first driving session, which is here considered as a training session to get familiar with driving in the simulator. From these 312 drives, 9310 one-minute segments were extracted (some time segments were excluded due to low signal quality), and for each of these observations, 275 features were calculated: 270 from EEG, 2 from EOG, and the remaining 3 features from contextual data.

3.2.1. Features from the EEG

3.2.1.1. EEG power spectra. The frequency characteristics of EEG signals are usually analysed using the power spectral density (PSD) in certain frequency bands, namely $\delta(<4\,\text{Hz})$, $\theta(4–7\,\text{Hz})$, $\alpha(8–12\,\text{Hz})$, $\beta(12–30\,\text{Hz})$, and γ (31–50 Hz). Here, from each EEG channel, PSDs were estimated with Welch's method (Welch, 1967) using a Blackman window with 50% overlap. In addition to the power in the α, β, θ and δ frequency bands, different ratios of these powers, $(\theta+\alpha)/\beta$, α/β, $(\theta+\alpha)/(\alpha+\beta)$, and θ/β (Jap, Lal, Fischer, & Bekiaris, 2009), were also used as features. This provided 9 features from each EEG channel, resulting in 270 EEG features from each 1-min segment.

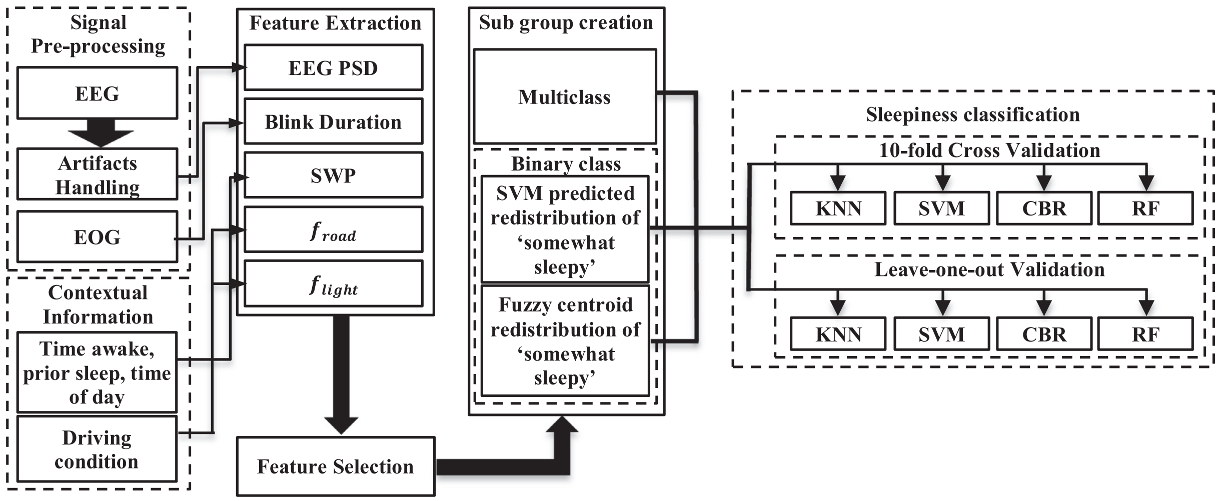

Fig. 3. Driver sleepiness classification scheme for experimental work.

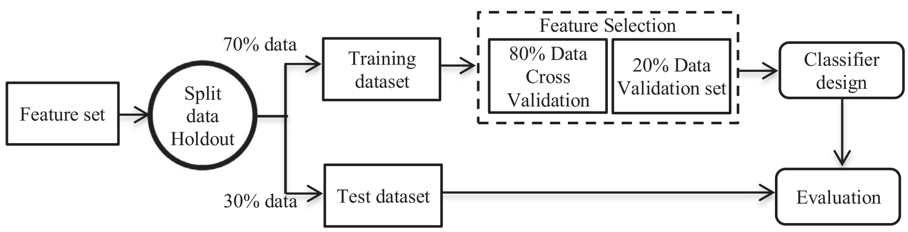

Fig. 4. Dataset preparation for feature selection and classification.

3.2.2. Features from the EOG

3.2.2.1. Blink duration. Blink durations were derived from the vertical EOG with an automatic blink detection algorithm based on derivatives and thresholding, see Jammes, Sharabty, and Esteve (2008) for details. The blink duration was calculated at half the amplitude of the upswing and downswing of each blink, respectively, to avoid issues with finding the exact onset and endpoint of the blink. The blink duration was then defined as the time duration between the two. The mean blink duration in each 1 min segment was used as a feature.

3.2.2.2. PERCLOS. The percentage of time that the eyelids were closed more than 80% (Wierwille, Wreggit, Kirn, Ellsworth, & Fairbanks, 1994), in each 1 min segment, was also extracted as a feature. The PERCLOS calculation was carried out as described by Jammes et al. (2008). To estimate PERCLOS, eyes wide open is considered as 0% and eyes closed is considered as 100%. PERCLOS has been found to be sensitive to sleep deprivation (Kozak et al., 2005) and to be a valid psychophysiological measure of alertness and sleepiness during driving (De Rosario, Solaz, Rodriguez, & Bergasa, 2010; Dinges & Grace, 1998).

3.2.3. Features from contextual information

3.2.3.1. Sleep/Wake predictor model (SWP). Sleepiness is mainly regulated by three factors, the circadian rhythm, time awake, and prior sleep. These variables can be used to construct mathematical models of sleep and sleepiness. Here, SWP was used since it has been shown to predict run-off-road crashes (Åkerstedt et al., 2008). SWP models the homeostatic effects of time awake and amount of prior sleep (S), the circadian effect that represents the influence of the biological clock (C), and an ultradian component (U). The level of alertness, formed as the sum $S+C+U$, is transformed into a level of sleepiness similar to KSS as $SWP=10.9 - 0.6(S+C+U)$. A detailed description of the model parameters and constants involved in calculating S, C, and U can be found in Sandberg et al. (2011). The SWP value obtained in the beginning of each 1 min segment was used as a feature.

3.2.3.2. Driving condition. Sleepiness is regulated by many different factors. In this study, the effect of two such factors were investigated; light conditions (simulated daylight versus simulated darkness) and driving environment (rural versus suburban). These factors were incorporated in the experimental design, where the drivers drove three different simulated scenarios: a rural road in simulated daylight, the same rural road but in simulated darkness, and a suburban road in simulated daylight. Two binary features were created to account for these two factors, see Eqs. (1) and (2). For observations recorded in simulated daylight, the light feature was assigned a 1, and for observations recorded in simulated darkness, this feature was assigned a 0. Similarly, the road environment feature was assigned a 0 when the driver drove the suburban scenario and a 1 when driving on the rural road.

$$f_{road} = \begin{cases} 1, & \textit{Rural road} \\ 0, & \textit{Suburban road} \end{cases}, \tag{1}$$

$$f_{light} = \begin{cases} 1, & \textit{Daylight in rural or suburban road} \\ 0, & \textit{Darkness in rural road} \end{cases}, \tag{2}$$

3.3. Feature selection

Since most of the features were extracted from the 30 EEG channels, many of which were recorded from neighbouring electrodes, we expected the extracted features to overlap. Hence, by reducing redundant and irrelevant features, the performance and generalisability of the classifiers can be improved. Three feature selection algorithms were assessed for deriving optimal feature subsets: a univariate feature selection method developed by

Dudoit, Fridlyand, and Speed (2002), a wrapper method called Sequential Forward Floating Selection (SFFS) (Mekyska et al., 2015; Pudil, 1994; Whitney, 1971), and a third method called minimum Redundancy Maximum Relevance (mRMR) (Hanchuan, Fuhui, & Ding, 2005). The univariate feature selection method was based on the ratio of the sum of squared differences between-groups and within-groups (BSS/WSS), for each feature, Eq. (3).

$$\frac{BSS(k)}{WSS(k)} = \frac{\sum_i \sum_j I_{i,j}\left(\bar{x}_{j,k} - \bar{x}_k\right)^2}{\sum_i \sum_j I_{i,k}\left(x_{i,k} - \bar{x}_{j,k}\right)^2}, \tag{3}$$

$x_{i,k}$ denotes the value of the k^{th} feature for the i^{th} observation, $\bar{x}_{j,k}$ is the average of the k^{th} feature that belongs to the class j, and $\bar{x}_k$ represents the average of the k^{th} feature over all observations. $I_{i,j}$ represents the parameter for each class according to Eq. (4):

$$I_{i,j} = I_{i,j} = \begin{cases} 1, & if\ i\ belongs\ to\ class\ j \\ 0, & otherwise \end{cases}, \tag{4}$$

Features were ranked so that higher rankings were given to the features that have large variation between classes and small variation within classes. BSS/WSS was chosen because of its simplicity and because features with higher discriminating power could be determined between classes. The algorithm is also computationally fast compared to other algorithms. Later, an SVM with a Gaussian kernel (see Section 3.4) was used to evaluate the performance of increasingly larger feature subsets by selecting features with BSS/WSS greater than a threshold value. Threshold values ranging between [0, 0.1] were used for selecting the feature subsets.

SFFS was used to also investigate the intra-feature relationships. SFFS is a successor of the sequential forward selection (SFS) method, which does not suffer from the "nesting effect", and which is computationally more efficient than other branch and bound methods (Pudil, 1994). SFFS performs a bottom-up search i.e., the algorithm starts with the null feature set, and in each step the best feature is added to the current feature set, followed by exclusion of the worst feature in the updated set (Chandrashekar & Sahin, 2014; Pudil, 1994). SFFS dynamically increase and decrease the number of features until the optimal feature subset is found. SFFS was wrapped with an SVM classifier with the same configuration as was used for BSS/WSS to select and evaluate the optimal feature subset.

Filter type methods such as BSS/WSS often result in high-dimensional redundant feature sets where many of the selected features contain very similar information. Wrapper type methods, such as SFFS, often result in low-dimensional non-redundant feature sets, that may be more sensitive to noisy data and sometimes do not generalize well. mRMR can be seen as a compromise, where features are selected to be maximally relevant but under the constraint that they should also be minimally redundant (Ding & Peng, 2003). Minimum redundancy was achieved by minimizing the mutual information between features, and maximum relevance was achieved by maximizing the mutual information between the features and the target class. Details of mRMR can be found in Hanchuan et al. (2005). Like BSS/WSS and SFFS, the mRMR was wrapped with an SVM to evaluate the optimal feature subset.

Only training data were used in the feature selection step. The training set was further divided into two sets, where 80% of the training data were used for BSS/WSS, SFFS, and mRMR, and 20% of the training data were used as a validation set, as shown in Fig. 4. All three feature selection methods used an SVM classifier (see Section 3.4), trained using 10-fold cross-validation.

3.4. Driver sleepiness classification

The sleepiness classification scheme was carried out using KSS as response variable. Three different subgroups were used as target values when training the classifiers: alert (KSS 1 – 5), somewhat sleepy ($6 \leq \text{KSS} \leq 7$) and sleepy ($\text{KSS} \geq 8$). Using subjective sleepiness ratings as the ground truth for sleepiness is motivated by the fact that KSS is easily applied, unobtrusive, commonly used in driver sleepiness research (Fu, Wang, & Zhao, 2016; Hu, Zheng, & Peters, 2013; Kong, Zhou, Jiang, Babiloni, & Borghini, 2017; Mu, Hu, & Min, 2017; Sandberg et al., 2011) and above all, the measure of driver sleepiness that is least affected by interindividual variations (Åkerstedt et al., 2013; Åkerstedt, Anund, Axelsson, & Kecklund, 2014). Other ground truths of driver sleepiness that have been used for driver sleepiness classification include expert ratings based on video recordings (Khushaba, Kodagoda, Lal, & Dissanayake, 2011; Li & Chung, 2015), expert ratings based on physiological signals (Picot, Charbonnier, & Caplier, 2012), the supposed alertness level that follows from an experimental design with sleep deprived participants (Jiřina, Bouchner, & Novotný, 2010), the percentage of eye closure (Li, Lee, & Chung, 2015), and lane departure events. However, video based expert ratings have been found to be unreliable (Ahlstrom, Fors, Anund, & Hallvig, 2015), the experimental design approach does not guarantee that the driver is alert in the supposedly alert condition, and lane departure events are rare in themselves and only reflects the rather rare lapses in attention that follows from insufficient sleep (so called wake state instability) (Doran, Van Dongen, & Dinges, 2001). Electrophysiological measures of sleepiness could not be used since they are used as features in the developed classifiers.

Classifiers were developed in three different manners: *multiclass* – considering all three groups, *binary class with SVM prediction*-based redistribution of the observations from the ‘somewhat sleepy’ group to the ‘alert’ and ‘sleepy’ groups, and *binary class with fuzzy centroid*-based redistribution of the observations from the ‘somewhat sleepy’ group to the ‘alert’ and ‘sleepy’ groups. The definition of the ‘sleepy’ group is motivated by the fact that KSS levels 8 and 9 are associated with severe signs of physiological sleepiness and increased frequency of lane-departure incidents (Reyner & Horne, 1998). Redistributing the ‘somewhat sleepy’ group to the crisper and more well-defined groups ‘alert’ and ‘sleepy’ is motivated by the fact that it is difficult to self-rate this “in-between” state (Sandberg et al., 2011). In SVM prediction-based redistribution, first an SVM classification model was created based on the training data set, where the ‘somewhat sleepy’ group was excluded from the training data. Then the SVM model was used to predict all the observations of the ‘somewhat sleepy’ group as ‘alert’ or ‘sleepy’ and assigned according to the obtained prediction. The fuzzy centroid (Dunn, 1973) was estimated from each of the ‘alert’ and ‘sleepy’ groups individually and then measuring the closest distance between ‘somewhat sleepy’ and the other two groups. If a ‘somewhat sleepy’ observation was closer to ‘alert’ than ‘sleepy’ then the observation was assigned to the alert group and vice versa.

In this study, the following four classifiers were compared:

KNN is a flexible and memory-based algorithm that does not require a model to be fit to the data and that uses the observations in the training set to find the most similar properties of the test dataset (Larose, 2005). The bias-variance trade-off of KNN depends of the selection of K, i.e., the number of nearest neighbours to be considered. Here, $K = 10$ was chosen with the intention to reduce the variance but limit the bias. The Euclidean distance function was used with a “squared inverse” distance weight. Moreover, the class with the nearest neighbour among tied groups was set for tie breaking.

SVM finds the hyperplane that not only minimizes the empirical classification error but also maximizes the geometric margin in the classification (Vapnik, 1992). SVM can map the original data points from the input space to a high dimensional feature space such that the classification problem becomes simple in this feature

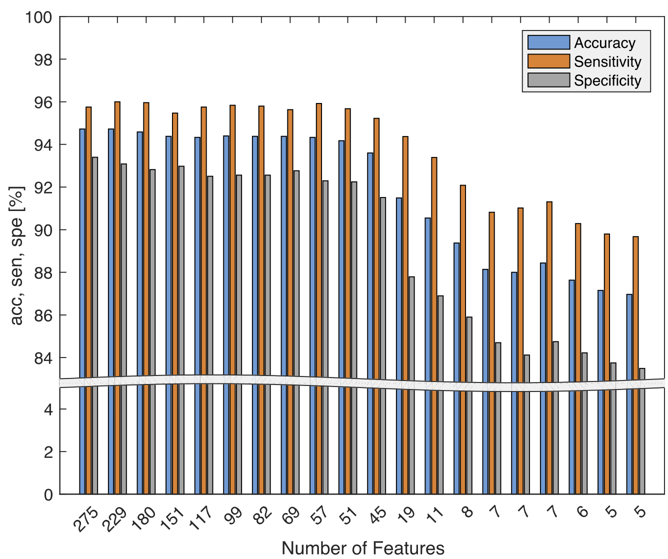

Fig. 5. Feature selection using BSS/WSS on the training dataset, validated using SVM and 10-fold cross-validation. Binary classification i.e., 'alert' and 'sleepy' groups were considered for the evaluation using SVM.

space. This makes the SVM suitable for classification problems with redundant data sets (Guyon, Weston, Barnhill, & Vapnik, 2002). In this study, an SVM with a Gaussian kernel was used for the classification task. The two SVM parameters C and gamma were tested and optimized within the range [0.001, 10], and in the final algorithm C was set to 0.03 and gamma to 10. Moreover, for multiclass classification, a one-against-all classification approach was considered in this paper.

CBR is a reasoner that solves a new problem by remembering and reusing previously solved problems (Aamodt & Plaza, 1994; Kolodner, 1992). In CBR, cases hold the raw knowledge of the problem domain and can be appropriate to use when the problem domain is poorly understood or defined. In this study, the training dataset was used as case-library. Similarities between the cases in the case-base and the target case were measured by the Euclidean distance function (details can be found in Begum, Ahmed, Funk, Xiong, and Von Schéele (2009)) and feature weights were estimated using the *Fisher criteria* as described in Yang, Kyrgyzov, Wiart, and Bloch (2013).

RF is an ensemble method that consists of series of randomized decision-trees, where the output is the majority vote of all these decision-trees (Breiman, 2001). One important aspect of RF is that it does not assume independence of features. In the driving context, data is often noisy and rarely linearly separable into sleepy or not (McDonald, Lee, Schwarz, & Brown, 2014). The RF was implemented using bagging as the ensemble method and the maximal number of decision splits used was 4357.

The developed classifiers were evaluated in terms of confusion matrices, Receiver Operating Characteristic (ROC) curves, accuracy, sensitivity, and specificity.

4. Results

4.1. Feature selection

Evaluations of BSS/WSS, SFFS, and mRMR for binary classification (excluding 'somewhat sleepy' group) are presented in Fig. 5, Fig. 6, and Fig. 7, respectively. For the BSS/WSS method, the classification accuracy reached the maximum value using 180 features (threshold value 0.02). Classification accuracy, sensitivity and specificity using all 275 features (threshold value 0) to using 57 features (threshold value 0.08) were almost the same as when using 180 features. However, with fewer features (less than 57), classification accuracy, sensitivity and specificity decreased. Hence, a feature subset with 57 features was selected from BSS/WSS as listed in Table 1. With SFFS, the best classification accuracy was about 90% when 10 features were used. The accuracy, sensitivity, specificity, and classification score (scr) are shown in Fig. 6. Scr measures the trade-off between sensitivity and specificity, defined as $2^{\sin(\frac{\pi.SEN}{2}).\sin(\frac{\pi.SPE}{2})}$ (Mekyska et al., 2015). Corresponding results for mRMR was an accuracy of 90% when 30 features were used, see Fig. 7.

All three feature selection procedures ranked SWP as the most important feature, see Table 1. The other two contextual features were ranked lower, with f_{light} ranked 43 and f_{road} ranked 47 in BSS/WSS. With SFFS, f_{light} was the fourth selected feature, and with mRMR, f_{light} ranked 28. Features according to BSS/WSS as well as mRMR essentially consist of clusters or groups containing similar information. For example, BSS/WSS included both the eye closure related EOG features, and also $(\theta+\alpha)/(\alpha+\beta)$ from eleven different frontal electrode sites. Similarly, mRMR included $(\theta+\alpha)/(\alpha+\beta)$ from four different frontal electrode sites. SFFS also selected $(\theta+\alpha)/(\alpha+\beta)$, but only from one electrode site. In Table 1, features from neighbouring and symmetric electrodes have been grouped together. Colour coding represents the feature groups that are similar in all feature selection methods and the superscript number represents the rank of each feature by each feature selection method.

Based on the performance on the validation dataset, it was found that an SVM classifier based on BSS/WSS generalised better than classifiers based on SFFS and mRMR. For BSS/WSS, the classification accuracy, sensitivity, and specificity were 94%, 96%, and 92% respectively, whereas the corresponding results were 86%, 92%, and 91% for SFFS and 90%, 86%, and 85% for mRMR. From here on, all results will therefore be based on the BSS/WSS feature set with 57 features. More detailed results based on the BSS/WSS subset can be found in Section 4.2.

Since the contextual features were of particular interest in this study, further comparisons were made when these features were included versus excluded from the feature sets. BSS/WSS was then used to select two feature subsets, with and without contextual features, followed by binary and multiclass SVM classification. The results are presented in Table 2. Both the best and the worst classification performance is presented, where best performance was achieved using a BSS/WSS threshold of 0.02 and worst performance was achieved with a threshold of 0.1 (using a threshold of 0.08 resulted in performance very similar to the 0.02 threshold). For both binary and multiclass classification, performance increased considerably when the contextual features were included, much because of the SWP feature.

4.2. Driver sleepiness classification

An overview of the classification results is provided in Fig. 8. As expected, in both the 10-fold cross-validation approach and in the LOO approach, binary classification scores were higher than multiclass classification scores. Decreased accuracy for multiclass and binary classification can be seen for all classifiers except CBR when LOO validation is used instead of 10-fold cross-validation. More detailed results are presented in Sections 4.2.1 and 4.2.2.

4.2.1. 10-fold cross-validation

4.2.1.1. Multiclass classification. Confusion matrices, including the number of correct classifications for the test dataset, are shown in Table 3. Over all, the SVM classifier showed best performance across the three target groups and across the different performance

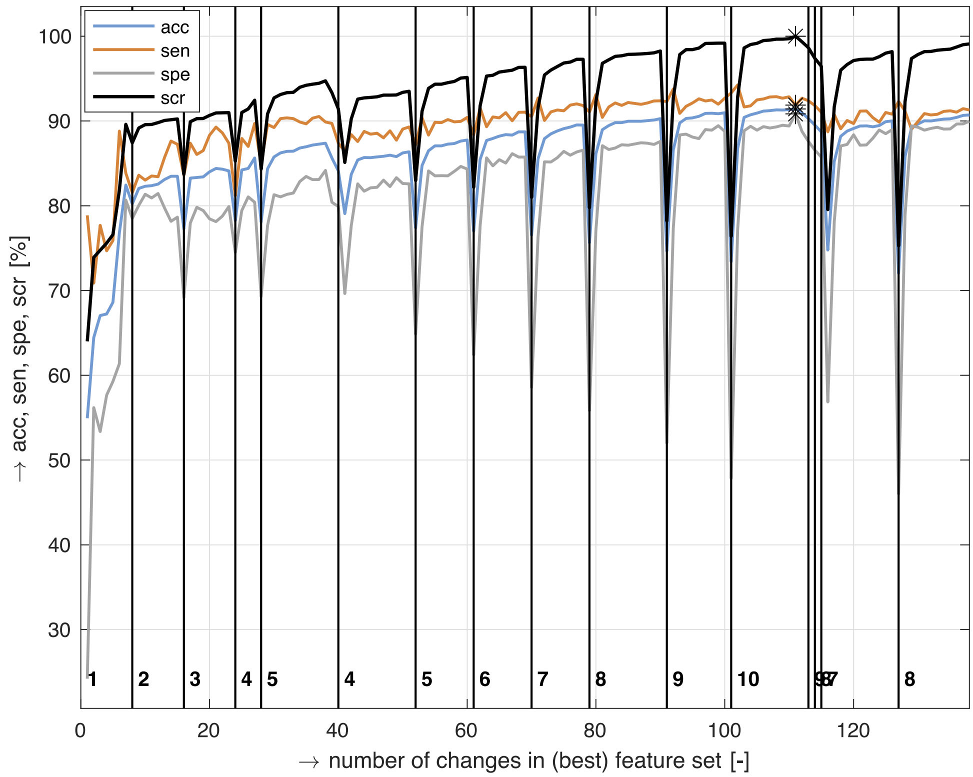

Fig. 6. Feature selection using SFFS on the training dataset, validated using SVM and 10-fold cross-validation. Similar to BSS/WSS, binary classification was considered for the evaluation using SVM.

Table 1
List of features included in the optimal feature set according to SFFS, BSS/WSS and mRMR. Features from neighbouring and symmetric electrodes are grouped together and colour coding shows the features that are similar between the three methods. The supercript number indicates the rank of the feature.

BSS/WSS	SFFS	mRMR
SWP^{1}	SWP^{1}	SWP^{1}
Blink duration^{2}	(θ+α)/β: F4^{2}	α: POz^{2}, O2^{14}, O1^{19}
PERCLOS^{3}	γ: FP2^{3}	(θ+α)/(α+β): FP2^{3}, FP1^{6}, FPz^{10}, F3^{12}
(θ+α)/(α+β): FP2^{4}, FP1^{5}, FPz^{6}, F3^{8}, F8^{12}, F7^{22}, F4^{41}, FC6^{42}, Fz^{46}, FC1^{52}, C4^{56}	f_light^{4}	Blink duration^{4}
γ: FC6^{7}, FP1^{32}, F3^{57}	θ: T8^{5}	θ/β: FPz^{5}, FP1^{13}, F7^{15}, FP2^{16}, FPz^{18}, F3^{26}
α/β: O2^{9}, O1^{11}, P4^{14}, POz^{15}, P8^{17}, Oz^{24}, Pz^{25}, P3^{40}	θ/β: T8^{6}	γ: O1^{7}
β: FC6^{10}, F3^{16}, F8^{19}, FP1^{33}, FC2^{39}, F4^{45}, F7^{49}	(θ+α)/(α+β): FPz^{7}	(θ+α)/β: F3^{8}, FP2^{23}
β: T7^{13}, T8^{18}, C4^{31}, CP5^{38}, P7^{44}, P4^{36}	α: FC5^{8}	θ: FP2^{9}, FP1^{21}, FPz^{30}
γ: T7^{20}, CP5^{28}, T8^{51}	α: O2^{9}	γ: P7^{11}, P8^{17}, T7^{27}
α/β: CP6^{21}, C4^{27}, FC6^{29}, CP2^{30}, FC2^{37}, F3^{48}, F4^{50}, Fz^{55}	α/β: P7^{10}	α/β: FP1^{20}, F3^{22}
α/β: T7^{23}, P7^{26}		β: Oz^{24}
β: Oz^{34}, O2^{53}, POz^{54}		α/β: P7^{25}
γ: Oz^{35}		f_light^{28}
f_light^{43}		α: FC2^{29}
f_road^{47}		

Table 2
Comparison of SVM evaluation of BSS/WSS feature selection. Results show best performance and worst performance when contextual features were included and excluded, respectively.

Criteria	Multiclass				Binary classification SVM prediction-based				Binary classification Fuzzy centroid-based			
	Including contextual features		Excluding contextual features		Including contextual features		Excluding contextual features		Including contextual features		Excluding contextual features	
	Best	Worst	Best	Worst	Best	Worst	Best	Worst	Best	Worst	Best	Worst
Accuracy	0.80	0.67	0.76	0.57	0.93	0.88	0.88	0.87	0.94	0.87	0.92	0.81
Sensitivity	0.85	0.74	0.80	0.61	0.92	0.91	0.91	0.90	0.94	0.88	0.91	0.85
Specificity	0.92	0.84	0.87	0.78	0.93	0.84	0.85	0.84	0.95	0.87	0.92	0.76

Table 3
Confusion matrix of KNN, SVM, CBR, and RF multiclass classification on the test dataset. The grey cells represent the true positive (TP) value. Here, TP represents the number of observations that were correctly classified, and the precision value in percentage.

Predicted class	Actual Class: KNN			SVM			CBR			RF		
	Alert	Somewhat sleepy	Sleepy	Alert	Somewhat sleepy	Sleepy	Alert	Somewhat sleepy	Sleepy	Alert	Somewhat sleepy	Sleepy
Alert	913 87%	118 11%	19 2%	893 85%	117 11%	40 4%	437 42%	309 29%	304 29%	917 87%	108 10%	25 3%
Somewhat sleepy	192 21%	636 69%	98 10%	147 16%	651 70%	128 14%	311 34%	318 34%	297 32%	211 23%	606 65%	109 12%
Sleepy	57 7%	138 17%	622 76%	40 5%	114 14%	663 81%	340 42%	260 32%	217 26%	50 6%	136 17%	631 77%

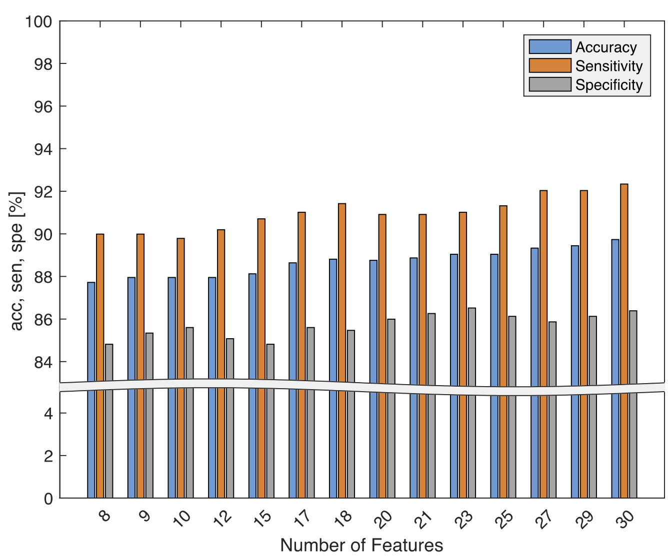

Fig. 7. Feature selection using mRMR on the training dataset, validated using SVM and 10-fold cross-validation. Binary classification i.e., 'alert' and 'sleepy' groups were considered for the evaluation using SVM.

measures. For the training dataset, the max, min, and average classification accuracy, within the 10-folds, were 81%, 76% and 78% for KNN, 83%, 77%, and 80% for SVM, 84%, 75%, and 77% for RF, and 36%, 30% and 33% for CBR.

The true positive (TP), true negative (TN), false positive (FP), false negative (FN), precision, sensitivity, specificity and balanced accuracy (BACC) on the test dataset is presented in Table 4. Here, *one-vs.-rest* was used to determine the KSS groups in the positive (P) and negative (N) classes, where the positive class is the KSS group that corresponds to each column (Alert, Somewhat sleepy and Sleepy) in Table 4, and the negative (N) class consists of other two KSS groups (Somewhat sleepy + Sleepy, Alert + Sleepy, and Alert + Somewhat sleepy). In terms of balanced accuracy, SVM performed better than the other three classifiers, with 87% for 'alert', 79% for 'somewhat sleepy' and 86% for 'sleepy'.

4.2.1.2. Binary classification. ROC curves for binary classification of the test dataset are illustrated in Fig. 9. The performance of KNN, SVM, and RF were similar and outperformed CBR. In general, the results for binary class classifications after SVM prediction-based redistribution of the 'somewhat sleepy' group were much better than classifications after fuzzy centroid based redistribution of the 'somewhat sleepy' group. TP, TN, FP, FN, sensitivity, specificity,

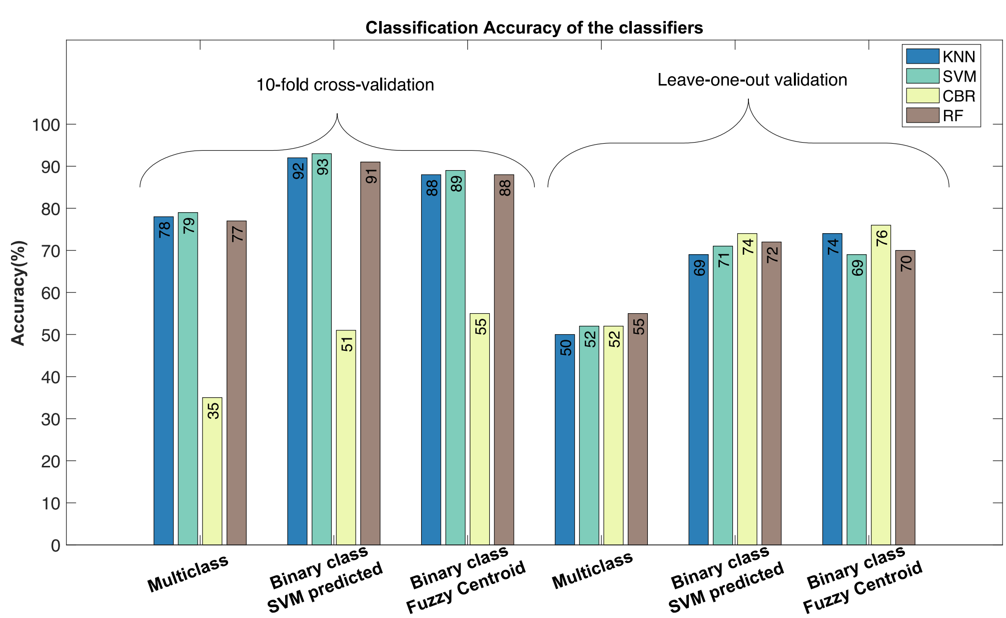

Fig. 8. Classification accuracy on the test dataset of KNN, SVM, CBR, and RF for multiclass, and two binary class classifications (redistributing the 'somewhat sleepy' group to the 'alert' and 'sleepy' groups using SVM based prediction and using fuzzy centroid). Two validation approaches, 10-fold cross-validation and LOO validation, are used.

Table 4
Performance summary of KNN, SVM, and CBR for multiclass classification on the test dataset.

Criteria	KNN			SVM			CBR			RF		
	Alert	Somewhat sleepy	Sleepy	Alert	Somewhat sleepy	Sleepy	Alert	Somewhat sleepy	Sleepy	Alert	Somewhat sleepy	Sleepy
KSS Group (P)	1050	926	817	1050	926	817	1050	926	817	1050	926	817
KSS Group (N)	1743	1867	1976	1743	1867	1976	1743	1867	1976	1743	1867	1976
TP	913	636	622	893	651	663	437	318	217	917	606	631
FP	137	290	195	157	275	154	613	608	600	133	320	186
FN	249	256	117	187	231	168	651	569	601	261	244	134
TN	1449	1622	1859	1556	1636	1808	1092	1298	1375	1482	1623	1842
Sensitivity	0.79	0.71	0.84	0.83	0.74	0.80	0.40	0.36	0.27	0.78	0.71	0.82
Specificity	0.92	0.85	0.91	0.91	0.86	0.92	0.64	0.68	0.70	0.92	0.84	0.91
Precision	0.86	0.80	0.89	0.88	0.82	0.88	0.55	0.58	0.57	0.86	0.80	0.89
BACC	0.86	0.77	0.85	0.87	0.79	0.86	0.52	0.52	0.48	0.86	0.76	0.85

Table 5
Performance summary of the classifiers for binary classification on the test dataset.

Criteria	SVM predicted redistribution of the "Somewhat Sleepy" Group				Fuzzy centroid redistribution of the "Somewhat Sleepy" Group			
	KNN	SVM	CBR	RF	KNN	SVM	CBR	RF
Total observations	2793	2793	2793	2793	2793	2793	2793	2793
Alert (P)	1491	1491	1491	1491	1679	1679	1679	1679
Sleepy (N)	1302	1302	1302	1302	1114	1114	1114	1114
TP	1400	1389	823	1379	1549	1533	1281	1529
FP	91	102	668	112	119	135	387	139
FN	138	82	705	143	217	177	859	206
TN	1164	1220	597	1159	908	948	266	919
Sensitivity	0.91	0.94	0.54	0.91	0.88	0.90	0.60	0.87
Specificity	0.93	0.92	0.47	0.91	0.88	0.88	0.41	0.88
Accuracy	0.92	0.93	0.51	0.91	0.88	0.89	0.55	0.88
BACC	0.92	0.93	0.50	0.91	0.87	0.88	0.50	0.87

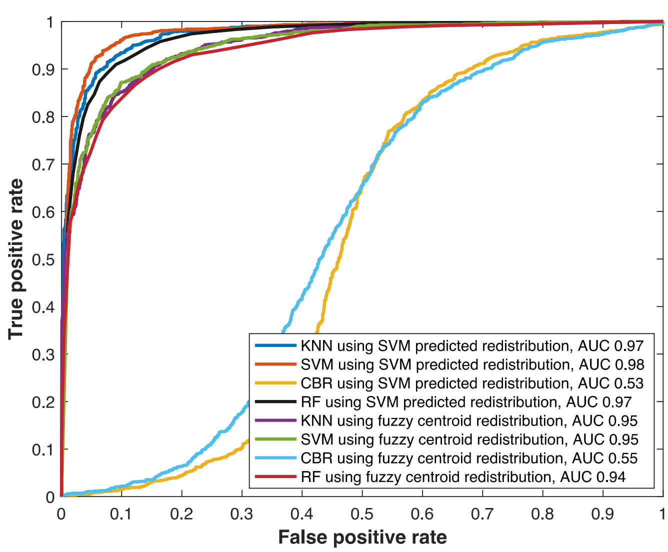

Fig. 9. ROC curves for KNN, SVM, CBR, and RF classification on the test dataset, where the models were trained using 10-fold cross-validation.

overall accuracy and balanced accuracy on the test dataset are presented in Table 5. On the training dataset, after SVM prediction-based redistribution of 'somewhat sleepy' group, the min, max and average classification accuracy across the 10 folds were 90%, 94% and 92% for KNN, 92%, 96%, and 94% for SVM, 45%, 53%, and 50% for CBR, and 90%, 92%, and 91% for RF. After Fuzzy centroid based redistribution of the 'somewhat sleepy' group , the min, max and average results were 86%, 90% and 88% for KNN, 89%, 91%, and 89% for SVM, 52%, 56%, and 54% for CBR, and 85%, 91%, and 88% for RF.

Compared to multiclass classification, the overall performance of KNN, SVM, and RF improved in binary classification. However, CBR didn't succeed in improving much in any of the categories.

4.2.2. Leave-one-out validation

4.2.2.1. Multiclass classification. confusion matrices showing the number of correct classifications on the test dataset are presented in Table 6. As was the case in 10-fold cross-validation, the 'somewhat sleepy' class is still problematic. A performance summary is provided in Table 7. RF shows slightly better results than the other three classifiers, with 72% for 'alert', 64% for 'somewhat sleepy' and 74% for 'sleepy'.

4.2.2.2. Binary classification. ROC curves showing test results from the LOO validation are provided in Fig. 10. Similar performance can be seen for KNN, SVM, and RF for both binary class categories. However, CBR performs better with Fuzzy centroid based redistribution than SVM prediction-based redistribution. The performance of the four classifiers on the test dataset is presented in Table 8.

LOO-investigation on the impact of individual differences: Corresponding results to Table 8, but where observations were left out from one KSS rating (5 min of data) instead of one participant, are provided in Table 9. Comparing Table 8 and Table 9, it can be seen that by incorporating data from the participant being tested into the training set, BACC increased with about 10% for all four classifiers.

5. Discussion

In this study, driver sleepiness detection based on EEG, EOG and contextual information is demonstrated using four well established classifiers. Amongst the contextual features, SWP was ranked as the most important feature in the entire dataset. The light condition feature also contributed substantially to increase the classification performance. The road environment does however not

Table 6
Confusion matrices of KNN, SVM, CBR, and RF multiclass classification. The grey cells represent the true positive value. Here, TP is represented the observation that were correctly classified and the precision value in percentage.

	Actual Class											
Predicted class	KNN			SVM			CBR			RF		
	Alert	Somewhat sleepy	Sleepy	Alert	Somewhat sleepy	Sleepy	Alert	Somewhat sleepy	Sleepy	Alert	Somewhat sleepy	Sleepy
Alert	2328 67%	834 24%	338 9%	2043 58%	881 25%	576 17%	2022 58%	1102 31%	376 11%	2382 68%	722 21%	396 11%
Somewhat sleepy	1120 36%	1273 42%	692 22%	887 29%	1294 42%	904 29%	971 31%	1334 44%	780 25%	1035 34%	1262 41%	788 25%
Sleepy	652 24%	982 30%	1091 56%	383 14%	820 30%	1522 56%	435 16%	847 31%	1443 53%	446 16%	796 29%	1483 55%

Table 7
Performance summary of KNN, SVM, CBR and RF for multiclass classification.

Criteria	KNN			SVM			CBR			RF		
	Alert	Somewhat sleepy	Sleepy	Alert	Somewhat sleepy	Sleepy	Alert	Somewhat sleepy	Sleepy	Alert	Somewhat sleepy	Sleepy
KSS Group (P)	1050	926	817	1050	926	817	1050	926	817	1050	926	817
KSS Group (N)	1743	1867	1976	1743	1867	1976	1743	1867	1976	1743	1867	1976
TP	2328	1273	1091	2043	1294	1522	2022	1334	1443	2382	1262	1483
FP	1172	1812	1624	1457	1791	1203	1478	1751	1282	1118	1823	1242
FN	1772	1816	1030	1270	1701	1480	1406	1949	1156	1481	1518	1184
TN	4038	4409	5555	4540	4524	5150	4404	4276	5429	4329	4707	5401
Sensitivity	0.57	0.41	0.51	0.62	0.43	0.51	0.59	0.41	0.56	0.62	0.45	0.56
Specificity	0.78	0.71	0.77	0.76	0.72	0.81	0.75	0.71	0.81	0.79	0.72	0.81
Accuracy	0.68	0.61	0.71	0.71	0.62	0.72	0.69	0.60	0.74	0.72	0.64	0.74
BACC	0.68	0.56	0.62	0.68	0.57	0.67	0.68	0.56	0.68	0.71	0.58	0.68

Table 8
Performance summary of the classifiers for binary classification.

Criteria	SVM predicted redistribution of “Somewhat Sleepy” Group				Fuzzy centroid redistribution of “Somewhat Sleepy” Group			
	KNN	SVM	CBR	RF	KNN	SVM	CBR	RF
Total observations	9310	9310	9310	9310	9310	9310	9310	9310
Alert (P)	3500	3500	3500	3500	5442	5442	5440	5442
Sleepy (N)	5810	5810	5810	5810	3868	3868	3870	3868
TP	2163	1878	2076	2073	4020	3667	4373	3855
FP	1337	1622	1424	1427	1420	1773	1067	1585
FN	1524	1063	1021	1157	1585	1067	1198	1244
TN	4286	4747	4789	4653	2285	2803	2672	2626
Sensitivity	0.59	0.64	0.67	0.64	0.72	0.77	0.78	0.76
Specificity	0.76	0.75	0.77	0.77	0.62	0.61	0.71	0.62
Accuracy	0.69	0.71	0.74	0.72	0.74	0.69	0.76	0.70
BACC	0.69	0.68	0.71	0.70	0.66	0.69	0.75	0.69

Table 9
Performance summary of the classifiers for binary classification considering participant dependent observations.

Criteria	SVM predicted redistribution of “Somewhat Sleepy” Group				Fuzzy centroid redistribution of “Somewhat Sleepy” Group			
	KNN	SVM	CBR	RF	KNN	SVM	CBR	RF
Total observations	9310	9310	9310	9310	9310	9310	9310	9310
Alert (P)	3500	3500	3500	3500	5442	5442	5440	5442
Sleepy (N)	5810	5810	5810	5810	3868	3868	3870	3868
TP	2303	2541	2498	2653	4231	4161	4706	4277
FP	897	959	1002	847	1209	1279	734	1163
FN	983	803	639	778	1118	905	837	990
TN	4827	5007	5171	5053	2752	2965	3033	2880
Sensitivity	0.73	0.76	0.80	0.77	0.79	0.82	0.85	0.81
Specificity	0.84	0.84	0.84	0.86	0.69	0.70	0.81	0.71
Accuracy	0.80	0.81	0.82	0.83	0.75	0.77	0.83	0.77
BACC	0.79	0.79	0.80	0.81	0.74	0.77	0.82	0.77

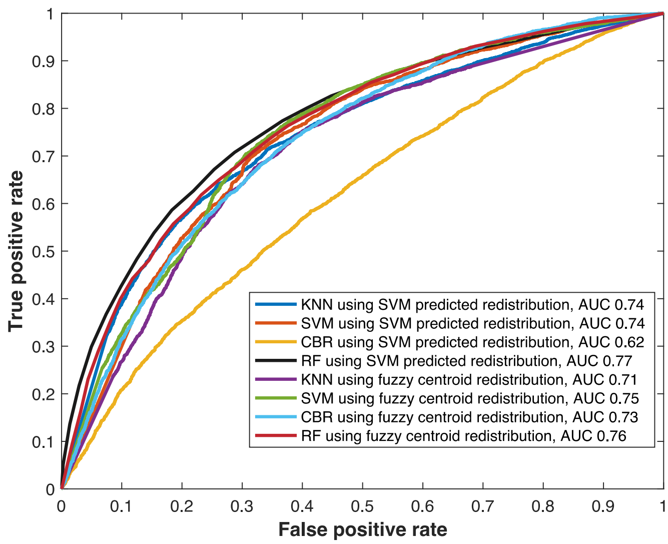

Fig. 10. ROC curves for KNN, SVM, CBR, and RF classification, where models were evaluated using leave-one-out validation.

seem to be that important for driver sleepiness classification. All in all, classification accuracy improved with 4% for multiclass classification and with 5% for binary classification when adding contextual features. Amongst the physiological features, blink duration derived from the EOG and different ratios between EEG power in the low frequency range (indicating sleepiness) and the high frequency range (indicating alertness) contributed most to the sleepiness classification results. Regarding the comparison of different classifiers, similar results were obtained for KNN, SVM and RF, whereas CBR performed worse. The choice of the classification algorithm thus plays a subordinate role compared to the choice of features, and also compared to whether a personalised algorithm can be used or not. When incorporating data from the participant being evaluated into the training set, classification accuracy increased by about 10% for all tested classifiers.

5.1. Feature selection

One of the objectives of this study was to identify a robust feature set. To obtain this feature set, three different features selection methods were evaluated: BSS/WSS, SFFS and mRMR. These three methods use different approaches to select the features. BSS/WSS seeks variation between classes with respect to the features and rank the features individually according to the ratio, whereas SFFS seeks intra-feature relationships between features. SFFS reduced the EEG features to a very small number, which indicates large overlap among the EEG features. This makes sense since many EEG features are extracted from neighbouring electrode sites. However, restricting the feature set to a small number of EEG features might be error-prone, which was seen in the classification results presented in Section 4.1. mRMR can be seen as a compromise, providing a more compact feature set than BSS/WSS, but with better generalizability compared to SFFS.

When grouping together EEG features from neighbouring or symmetric electrode sites (Table 1), it was noticed that the number of selected feature groups were very similar for all three algorithms. The number of features per group was largest for BSS/WSS, followed by mRMR. For SFFS, there was only one feature per group, indicating that there is very little redundancy in the SFFS feature set. However, some redundancy may be a good thing, especially in a noisy dataset such as the driver sleepiness dataset in this study. This line of reasoning explains the evaluation results, which shows that the feature subset selected with BSS/WSS achieved higher accuracy than the other two algorithms, 8% higher than SFFS and 4% higher than mRMR. It is worth mentioning that despite the many seemingly redundant features selected by BSS/WSS, the feature subset was still reduced to almost one fifth of the original feature set.

Regardless of the feature selection approach used, SWP always achieved the highest rank. SWP has previously been found to be a valuable driver sleepiness indicator (Mårtensson et al., 2018; Sandberg et al., 2011), especially on a population level (Åkerstedt et al., 2008). The finding of SWP as a top-ranking indicator suggests that it can also be used for individualised sleepiness detection when combined with other features. The light condition feature was also selected by all three feature selection algorithms, notably as the fourth included feature by SFFS (Table 1). This finding is also supported by the literature, where it has previously been found that darkness has an additive effect on driver sleepiness (Ahlström, Anund, Fors, & Åkerstedt, 2018a). The final contextual feature, the road environment, did however not contribute much to driver sleepiness classification.

With BSS/WSS, both EOG features ranked higher than any of the EEG features (Table 1), and with mRMR, blink duration was ranked as number 4. This is not surprising as PERCLOS is characteristic of drowsiness and correlates with vigilance (Dinges & Grace, 1998), whereas increased blink durations has been found for increasing levels of driver sleepiness in essentially all studies were blink duration has been measured (e.g., Åkerstedt et al. (2005), Schleicher et al. (2008)). Regarding the EEG features, the most dominating features are the ones that account for the energy ratio between lower frequencies (α and θ, indicating sleepiness) and higher frequencies (especially β, indicating alertness). This result agrees with Jap et al. (2009). The most dominating electrode sites are in the frontal areas as well as in the occipital/parietal areas. This also makes sense given the role of the frontal lobe in tasks related to attention, short-term memory and planning, and the role of the occipital lobe in visual information processing. From a practical point of view, future studies should identify the most relevant EEG electrodes in order to reduce the complexity in the sensor setup while maintaining classification performance. It would be interesting to investigate how much performance deteriorates when only using data from one or a few neighbouring electrode sites. For example, Rohit et al. (2017) only used features from frontal EEG channels to classify drowsiness in real-time, with similar results as we obtained in Table 1.

5.2. Classification

KNN, SVM and RF all showed similar classification performance across the different evaluations. SVM had the highest AUC in 10-fold cross-validation while RF had the highest AUC in LOO, Figs. 9 and 10. Although higher classification accuracy was obtained using CBR in LOO, RF was more stable (i.e. had higher probability of ranking a randomly chosen positive observation than ranking a randomly chosen negative observation). Despite the small overall differences between KNN, SVM and RF, we still recommend SVM since it performs well across all evaluations and was found to be the most stable classifier. This choice is also supported by Balandong et al. (2018), who reports that SVM is the most commonly used driver sleepiness classifier.

In machine learning it is important to design classifiers that generalize to unseen data, and the cornerstone to achieve this is proper validation of the algorithms. In real-time systems, generalization is highly relevant. To our knowledge, most developed ADSD systems have been evaluated using cross validation, whereas subject independent validation has rarely been used. Hence, two complementary evaluation approaches were used in this study, leave one participant out and 10-fold cross validation. Classification accuracy on the test dataset decreased for KNN, SVM, RF and increased for CBR in LOO validation compared to 10-fold cross-validation (Fig. 8). This result suggest that stability (Evgeniou, Pontil, & Elisseeff, 2004) of KNN, SVM, and RF is obtained when using 10-fold cross-validation but not when using LOO, especially when LOO refers to subject-independent evaluation. When including data from the participant being evaluated in the LOO training dataset, the balanced accuracy, sensitivity, and specificity increased with about 10% (Tables 8 and 9), turning the LOO results more in line with the 10-fold cross validation results. This highlights the added value of personalised algorithms, or at least generalized algorithms that can adapt to individuals over time. Generalizing a sleepiness detection model is a major challenge because of inter- and intra-individual variability (Jacobé de Naurois et al., 2017; Wang & Xu, 2016), meaning that physiological signals varies both between individuals and over time also within individuals, leading to drowsiness profiles that evolves over time within individual drivers (for details see Jacobé de Naurois et al. (2017)). The achieved classification results, especially in the leave-one-subject-out case, are lower than desirable in an ADSD system. Since the results are in line with previously published works (Balandong et al., 2018; Fu et al., 2016; Jacobé de Naurois et al., 2017; Wang & Xu, 2016), this highlights the difficulty in classifying driver sleepiness with very high accuracy.

The somewhat contradictory results obtained with CBR, with increasing performance in LOO compared to 10-fold cross validation, may seem strange at first. However, since CBR uses historical cases to solve a problem (Watson & Marir, 1994), it requires a sufficiently large case library to perform well. With LOO, the case library becomes bigger and thus covers a larger problem space, giving CBR a better likelihood of finding similar cases. An advantage with CBR is its inherent ability to continuously incorporate new cases into the case library, thus facilitating a system that adapts to the individual over time.

One observation from the multiclass classifications, presented in Tables 3 and 6, is that it is more difficult to correctly classify the 'somewhat sleepy' group compared to the 'alert' and 'sleepy' groups. Our results from multiclass classification are lower compared to the results reported in literature. However, those studies (Khushaba, Kodagoda, Lal, & Dissanayake, 2013; Lee, Lee, & Chung, 2014; Zhang, Wang, & Fu, 2014; Zhao, Zheng, Zhao, Tu, & Liu, 2011) used expert ratings based on video recordings for sleepiness classification, a target value that has been found to be unreliable. However, although KSS is a validated method for measuring sleepiness (Åkerstedt et al., 2014), there are difficulties in rating sleepiness in the transitional state between alert and sleepy. This issue with uncertainty in the target values may be alleviated by redistributing the 'somewhat sleepy' group into the 'alert' and 'sleepy' groups, which indeed led to improved performance. A noteworthy observation is that SVM prediction-based redistribution redistributed the 'somewhat sleepy' group more towards the 'sleepy' group, whereas the opposite was found for the fuzzy centroid based redistribution, see Table 8. Similar classification results were achieved with both binary categories; however, it would be interesting to investigate this further with a larger dataset.

There are a number of limitations to the ADSD suggested in this study. *First*, the physiological sensors that are used are both obtrusive and inconvenient for the driver. Many studies have therefore suggested vision based ADSD systems (Assari & Rahmati, 2011; Buciu, Gacsádi, & Grava, 2010; Flores, Armingol, & de la Escalera, 2010). However, camera-based systems are still struggling with occlusion, light conditions, focusing on small regions of interest, extracting multiple indicators from single cameras, etc. (Buciu et al., 2010; Ming-Hsuan, Kriegman, & Ahuja, 2002). Non-intrusive sensors and non-contact systems for driver monitoring is currently under development and in a near future unobtrusive sensors have hopefully matured to a level where they can be used in real-world applications such as driving (Rahman, Barua, & Shahina, 2015; Zheng et al., 2014). For example, ECG can be measured using sensors in the steering wheel, in the seat or in the seatbelt (Pinto, Cardoso, Lourenço, & Carreiras, 2017); camera-based solutions can be used to measure heart rate, heart rate variability (Rahman, Ahmed, & Begum, 2016) and eye closures (Golz et al., 2010), and EEG can be measured using dry electrodes on the scalp or in-ear EEG sensors (Fiedler, Obleser, Lunner, & Graversen, 2016; Hwang, Kim, Hong, & Park, 2016). The implementation of such systems in vehicles is also progressing, for example, Kartsch, Benatti, Schiavone, Rossi, and Benini (2018) have developed a parallel ultra-low power platform for detecting driver drowsiness that is comprised of both physiological signals such as EEG, EOG, and behavioural information from an inertial measurement unit. *Second*, there are challenges involved in incorporating the contextual features in real life applications. However, modern vehicles are already equipped with wireless networks and can synchronise with smart devices. In the future, personalised information such sleeping hours and time awake can be integrated via wearables and synchronised to the vehicle. Ambient light sensors are readily available in modern vehicles and can thus provide the information about light conditions, and companies such as *TomTom* and *Here* are currently working on integrating continuously updated traffic complexity information into their digital maps. Contextual features can also be computed as prior probabilities as proposed by Fu et al. (2016). Regarding SWP, the underlying biomathematical sleepiness model can be simplified to only considering driving time and time of day, and Mårtensson et al. (2018) showed promising results by just assuming that the person got up at 7 am. *Third,* measuring EEG in a driving context is problematic. Artifacts handling can improve the classification accuracy, however, using EEG features alone may not be enough since in real-world driving, the magnitude of changes in EEG signals have been shown to not correlate well with cognitive impairment (Sparrow et al., 2018). Sparrow et al. (2018) therefore suggested that EEG should not be used as the only source of information in driver sleepiness detection systems. In support of this claim, it has been put forward that unimodal features are not reliable enough (Yang, Lin, & Bhattacharya, 2005) and that multimodal information is needed (Jacobé de Naurois et al., 2017; Kartsch et al., 2018) for accurate sleepiness detection. The more complex sensor setup is obviously problematic, but it may be unavoidable given how notoriously difficult it is to measure sleepiness in a reliable manner (Mullington et al., 2011).

5.3. Conclusions

In conclusion, this paper has presented an ADSD system based on EEG, EOG, and contextual information. The multiclass classification results show that it is difficult to correctly classify the 'somewhat sleepy' group, but results improve when the 'somewhat sleepy' group is reassigned to the 'alert' and 'sleepy' groups, turning the problem into a binary classification task. Further, large improvements were observed when data from the driver being evaluated was included in the training set (about 10% increase in classification accuracy). The added value of the contextual features was an improved classification accuracy of 4% for multiclass classification and with 5% for binary classification. Future research should investigate (a) how more contextual information can be included in the system, and (b), how classifiers should be designed to continuously develop and adapt to the current driver as new unlabelled data becomes available.

Acknowledgment

The authors would like to acknowledge VINNOVA (Swedish Governmental Agency for Innovation Systems) for supporting the Vehicle Driver Monitoring (VDM) research project. The authors would also like to acknowledge our project partners Emma Nilsson, Per Lindén and Bo Svanberg (Volvo Car Corporation) as well as Anna Anund and Carina Fors (VTI).

References

Aamodt, A., & Plaza, E. (1994). Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI Communications, 7*, 39–59.
Abeyratne, U. R., Vinayak, S., Hukins, C., & Duce, B. (2009). A new measure to quantify sleepiness using higher order statistical analysis of EEG. In *Proceedings of the annual international conference of the IEEE engineering in medicine and biology society* (pp. 5543–5546).
Aeschbach, D., Matthews, J. R., Postolache, T. T., Jackson, M. A., Giesen, H. A., & Wehr, T. A. (1997). Dynamics of the human EEG during prolonged wakefulness: Evidence for frequency-specific circadian and homeostatic influences. *Neuroscience Letters, 239*, 121–124.
Ahlström, C., Anund, A., Fors, C., & Åkerstedt, T. (2018a). The effect of daylight versus darkness on driver sleepiness: A driving simulator study. *Journal of Sleep Research, 27*, e12642.
Ahlström, C., Anund, A., Fors, C., & Åkerstedt, T. (2018b). Effects of the road environment on the development of driver sleepiness in young male drivers. *Accident Analysis & Prevention, 112*, 127–134.
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.
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*, 242–254.
Åkerstedt, T., Connor, J., Gray, A., & Kecklund, G. (2008). Predicting road crashes from a mathematical model of alertness regulation—The Sleep/Wake Predictor. *Accident Analysis & Prevention, 40*, 1480–1485.
Åkerstedt, T., & Gillberg, M. (1990). Subjective and objective sleepiness in the active individual. *International Journal of Neuroscience, 52*, 29–37.
Åkerstedt, T., Hallvig, D., Anund, A., Fors, C., Schwarz, J., & Kecklund, G. (2013). Having to stop driving at night because of dangerous sleepiness – awareness, physiology and behaviour. *Journal of Sleep Research, 22*, 380–388.
Åkerstedt, T., Knutsson, A., Westerholm, P., Theorell, T., Alfredsson, L., & Kecklund, G. (2004). Mental fatigue, work and sleep. *Journal of Psychosomatic Research, 57*, 427–433.
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.
Akin, M., Kurt, M. B., Sezgin, N., & Bayram, M. (2008). Estimating vigilance level by using EEG and EMG signals. *Neural Computing and Applications, 17*, 227–236.
Anderson, C., & Horne, J. A. (2013). Driving drowsy also worsens driver distraction. *Sleep Medicine, 14*, 466–468.
Anund, A., Fors, C., Hallvig, D., Åkerstedt, T., & Kecklund, G. (2013). Observer rated sleepiness and real road driving: An explorative study. *PLoS ONE, 8*, e64782.
Assari, M. A., & Rahmati, M. (2011). Driver drowsiness detection using face expression recognition. In *Proceedings of the IEEE International Conference on Signal and Image Processing Applications (ICSIPA)* (pp. 337–341).
Babaeian, M., Bhardwaj, N., Esquivel, B., & Mozumdar, M. (2016). Real time driver drowsiness detection using a logistic-regression-based machine learning algorithm. In *Proceedings of the IEEE Green Energy and Systems Conference (IGSEC)* (pp. 1–6).
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., Ahlstrom, C., Begum, S., & Funk, P. (2017). Automated EEG artifact handling with application in driver monitoring. *IEEE Journal of Biomedical and Health Informatics*, 1 –1.
Begum, S., Ahmed, M. U., Funk, P., & Filla, R. (2012). Mental state monitoring system for the professional drivers based on heart rate variability analysis and case-based reasoning. In *Proceedings of the Federated Conference on Computer Science and Information Systems (FedCSIS)* (pp. 35–42).
Begum, S., Ahmed, M. U., Funk, P., Xiong, N., & Von Schéele, B. (2009). A case-based decision support system for individual stress diagnosis using fuzzy similarity matchING. *Computational Intelligence, 25*, 180–195.
Begum, S., Barua, S., Filla, R., & Ahmed, M. U. (2014). Classification of physiological signals for wheel loader operators using multi-scale entropy analysis and case-based reasoning. *Expert Systems with Applications, 41*, 295–305.
Boostani, R., Karimzadeh, F., & Nami, M. (2017). A comparative review on sleep stage classification methods in patients and healthy individuals. *Computer Methods and Programs in Biomedicine, 140*, 77–91.
Borghini, G., Astolfi, L., Vecchiato, G., Mattia, D., & Babiloni, F. (2014). Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. *Neuroscience & Biobehavioral Reviews, 44*, 58–75.
Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32.
Buciu, I., Gacsádi, A., & Grava, C. (2010). Vision based approaches for driver assistance systems. In *Proceedings of the eleventh WSEAS international conference on automation & information* (pp. 92–97). World Scientific and Engineering Academy and Society (WSEAS).
Cajochen, C., Brunner, D. P., Kräuchi, K., Graw, P., & Wirz-Justice, A. (1995). Power density in theta/alpha frequencies of the waking EEG progressively increases during sustained wakefulness. *Sleep, 18*, 890–894.
Chacon-Murguia, M. I., & Prieto-Resendiz, C. (2015). Detecting driver drowsiness: A survey of system designs and technology. *IEEE Consumer Electronics Magazine, 4*, 107–119.
Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering, 40*, 16–28.
Chen, L.-l., Zhao, Y., Ye, P.-f., Zhang, J., & Zou, J.-z. (2017). Detecting driving stress in physiological signals based on multimodal feature analysis and kernel classifiers. *Expert Systems with Applications, 85*, 279–291.
Chui, K. T., Tsang, K. F., Chi, H. R., Wu, C. K., & Ling, B. W. K. (2015). Electrocardiogram based classifier for driver drowsiness detection. In *Proceedings of the IEEE thirteenth international conference on industrial informatics (INDIN)* (pp. 600–603).
Čolić, A., Marques, O., & Furht, B. (2014). Driver Drowsiness Detection and Measurement Methods. In *Driver drowsiness detection: systems and solutions* (pp. 7–18). Springer International Publishing.
Cooke, J. R., & Ancoli-Israel, S. (2011). Chapter 41- Normal and abnormal sleep in the elderly. In P. Montagna, & S. Chokroverty (Eds.). In *Handbook of clinical neurology: 98* (pp. 653–665). Elsevier.
Craig, A., Tran, Y., Wijesuriya, N., & Nguyen, H. (2012). Regional brain wave activity changes associated with fatigue. *Psychophysiology, 49*, 574–582.
Curcio, G., Casagrande, M., & Bertini, M. (2001). Sleepiness: Evaluating and quantifying methods. *International Journal of Psychophysiology, 41*, 251–263.
De Rosario, H., Solaz, J. S., Rodriguez, N., & Bergasa, L. M. (2010). Controlled inducement and measurement of drowsiness in a driving simulator. *Intelligent Transport Systems, IET, 4*, 280–288.
Ding, C., & Peng, H. (2003). Minimum redundancy feature selection from microarray gene expression data. In *Proceedings of the IEEE bioinformatics conference computational systems bioinformatics CSB2003* (pp. 523–528).
Dinges, D. F. (1995). An overview of sleepiness and accidents. *Journal of Sleep Research, 4*, 4–14.
Dinges, D. F., & Grace, R. (1998). PERCLOS: A valid psychophysiological measure of alertness as assessed by psychomotor vigilance. *US Department of Transportation, Federal Highway Administration, Publication Number FHWA-MCRT, 98*, 006.
Doran, S., Van Dongen, H., & Dinges, D. F. (2001). Sustained attention performance during sleep deprivation: Evidence of state instability. *Archives italiennes de biologie, 139*, 253–267.
Dudoit, S., Fridlyand, J., & Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association, 97*, 77–87.
Dunn, J. C. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybernetics, 3*, 32–57.
Dwivedi, K., Biswaranjan, K., & Sethi, A. (2014). Drowsy driver detection using representation learning. In *Proceedings of the IEEE International Advance Computing Conference (IACC)* (pp. 995–999).
Elisseeff, A., & Pontil, M. (2003). Leave-one-out error and stability of learning algorithms with applications. In J. Suykens, G. Horvath, S. Basu, C. Micchelli, & J. Vandewalle (Eds.). *Learning theory and practice*: 190. IOS Press.
Evgeniou, T., Pontil, M., & Elisseeff, A. (2004). Leave one out error, stability, and generalization of voting combinations of classifiers. *Machine Learning, 55*, 71–97.
Fiedler, L., Obleser, J., Lunner, T., & Graversen, C. (2016). Ear-EEG allows extraction of neural responses in challenging listening scenarios – A future technology for hearing aids? In *Proceedings of the thirty-eight annual international conference of the IEEE engineering in medicine and biology society (EMBC)* (pp. 5697–5700).
Figueiro, M. G., Bullough, J. D., & Rea, M. S. (2007). *Light isn't just for vision anymore: implications for transportation safety*. Lighting Research Center, Rensselaer Polytechnic Institute.
Flores, M. J., Armingol, J. M., & de la Escalera, A. (2010). Real-time warning system for driver drowsiness detection using visual information. *Journal of Intelligent & Robotic Systems, 59*, 103–125.
Fors, C., Ahlstrom, C., & Anund, A. (2018). A comparison of driver sleepiness in the simulator and on the real road. *Journal of Transportation Safety & 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.
Garcés Correa, A., Orosco, L., & Laciar, E. (2014). Automatic detection of drowsiness in EEG records based on multimodal analysis. *Medical Engineering & Physics, 36*, 244–249.
Golz, M., Sommer, D., Trutschel, U., Sirois, B., & Edwards, D. (2010). Evaluation of fatigue monitoring technologies. *Somnologie - Schlafforschung und Schlafmedizin, 14*, 187–199.
Gonçalves, M., Amici, R., Lucas, R., Åkerstedt, T., Cirignotta, F., Horne, J., National Representatives as Study, C., et al. (2015). Sleepiness at the wheel across Europe: A survey of 19 countries. *Journal of Sleep Research, 24*, 242–253.
Gurudath, N., & Riley, H. B. (2014). Drowsy driving detection by EEG analysis using wavelet transform and k-means clustering. *Procedia Computer Science, 34*, 400–409.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning, 46*, 389–422.

Häkkänen, H., LicPsych, Summala, H., Partinen, M., Tiihonen, M., & Silvo, J. (1999). Blink duration as an indicator of driver sleepiness in professional bus drivers. *Sleep, 22*, 798–802.

Hallvig, D., Anund, A., Fors, C., Kecklund, G., Karlsson, J. G., Wande, M., et al. (2013). Sleepy driving on the real road and in the simulator–A comparison. *Accident Analysis and Prevention, 50*, 44–50.

Hanchuan, P., Fuhui, L., & Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27*, 1226–1238.

Hassan, A. R., & Bhuiyan, M. I. H. (2016a). Computer-aided sleep staging using complete ensemble empirical mode decomposition with adaptive noise and bootstrap aggregating. *Biomedical Signal Processing and Control, 24*, 1–10.

Hassan, A. R., & Bhuiyan, M. I. H. (2016b). A decision support system for automatic sleep staging from EEG signals using tunable Q-factor wavelet transform and spectral features. *Journal of Neuroscience Methods, 271*, 107–118.

Hassan, A. R., & Bhuiyan, M. I. H. (2017). An automated method for sleep staging from EEG signals using normal inverse Gaussian parameters and adaptive boosting. *Neurocomputing, 219*, 76–87.

Horne, J., & Reyner, L. (1999). Vehicle accidents related to sleep: A review. *Occupational and Environmental Medicine, 56*, 289–294.

Hu, S., & Zheng, G. (2009). Driver drowsiness detection with eyelid related parameters by support vector machine. *Expert Systems with Applications, 36*, 7651–7658.

Hu, S., Zheng, G., & Peters, B. (2013). Driver fatigue detection from electroencephalogram spectrum after electrooculography artefact removal. *IET Intelligent Transport Systems, 7*, 105–113.

Hwang, T., Kim, M., Hong, S., & Park, K. S. (2016). Driver drowsiness detection using the in-ear EEG. In *Proceedings of the thirty-eightth annual international conference of the IEEE engineering in medicine and biology society (EMBC)* (pp. 4646–4649).

ITF. (2017). *Road safety annual report 2017*. OECD Publishing.

Jacobé de Naurois, C., Bourdin, C., Stratulat, A., Diaz, E., & Vercher, J.-L. (2017). Detection and prediction of driver drowsiness using artificial neural network models. *Accident Analysis & Prevention*.

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*(3), 227–232.

Jap, B. T., Lal, S., Fischer, P., & Bekiaris, E. (2009). Using EEG spectral components to assess algorithms for detecting fatigue. *Expert Systems with Applications, 36*, 2352–2359.

Jiao, Y., & Lu, B. L. (2017). Detecting driver sleepiness from EEG alpha wave during daytime driving. In *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 728–731).

Jiřina, M., Bouchner, P., & Novotný, S. (2010). Identification of driver's drowsiness using driving information and EEG. *Neural Network World, 20*, 773.

Kar, S., Bhagat, M., & Routray, A. (2010). EEG signal analysis for the assessment and quantification of driver's fatigue. *Transportation Research Part F: Traffic Psychology and Behaviour, 13*, 297–306.

Kartsch, V. J., Benatti, S., Schiavone, P. D., Rossi, D., & Benini, L. (2018). A sensor fusion approach for drowsiness detection in wearable ultra-low-power systems. *Information Fusion, 43*, 66–76.

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.

Khushaba, R. N., Kodagoda, S., Lal, S., & Dissanayake, G. (2013). Uncorrelated fuzzy neighborhood preserving analysis based feature projection for driver drowsiness recognition. *Fuzzy Sets and Systems, 221*, 90–111.

Kolodner, J. (1992). An introduction to case-based reasoning. *Artificial Intelligence Review, 6*, 3–34.

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. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 49*, 1917–1921.

Kurt, M. B., Sezgin, N., Akin, M., Kirbas, G., & Bayram, M. (2009). The ANN-based computing of drowsy level. *Expert Systems with Applications, 36*, 2534–2542.

Larose, D. T. (2005). k-Nearest neighbor algorithm. In *Discovering knowledge in data* (pp. 90–106). John Wiley & Sons, Inc.

Lee, B.-G., Lee, B.-L., & Chung, W.-Y. (2014). Mobile healthcare for automatic driving sleep-onset detection using wavelet-based EEG and respiration signals. *Sensors, 14*, 17915–17936.

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

Li, G., & Chung, W. Y. (2018). Combined EEG-gyroscope-tDCS brain machine interface system for early management of driver drowsiness. *IEEE Transactions on Human-Machine Systems, 48*, 50–62.

Li, G., Lee, B. L., & Chung, W. Y. (2015). Smartwatch-based wearable EEG system for driver drowsiness detection. *IEEE Sensors Journal, 15*, 7169–7180.

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.

Lyles, D. (2015). Traffic Fatalities Fall in 2014, But Early Estimates Show 2015 Trending Higher. *In (Vol. 2017)*. NHTSA.

Ma, J., Murphey, Y. L., & Zhao, H. (2015). Real time drowsiness detection based on lateral distance using wavelet transform and neural network. In *Proceedings of the IEEE Symposium Series on Computational Intelligence* (pp. 411–418).

Mårtensson, H., Keelan, O., & Ahlström, C. (2018). Driver sleepiness classification based on physiological data and driving performance from real road driving. *IEEE Transactions on Intelligent Transportation Systems*, 1–10.

McDonald, A. D., Lee, J. D., Schwarz, C., & Brown, T. L. (2014). Steering in a random forest: Ensemble learning for detecting drowsiness-related lane departures. *Human Factors, 56*, 986–998.

Mekyska, J., Galaz, Z., Mzourek, Z., Smekal, Z., Rektorova, I., Eliasova, I., et al. (2015). Assessing progress of Parkinson's disease using acoustic analysis of phonation. In *Proceedings of the fourth international work conference on bioinspired intelligence (IWOBI)* (pp. 111–118).

Menefee, L. A., Cohen, M. J. M., Anderson, W. R., Doghramji, K., Frank, E. D., & Lee, H. (2000). Sleep disturbance and nonmalignant chronic pain: A comprehensive review of the literature. *Pain Medicine, 1*, 156–172.

Ming-Hsuan, Y., Kriegman, D. J., & Ahuja, N. (2002). Detecting faces in images: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24*, 34–58.

Mu, Z., Hu, J., & Min, J. (2017). Driver fatigue detection system using electroencephalography signals based on combined entropy features. *Applied Sciences, 7*, 150.

Mullington, J. M., Czeisler, C. A., Goel, N., Krueger, J. M., Balkin, T. J., Johns, M., et al. (2011). Panel discussion: Current status of measuring sleepiness. *Journal of clinical sleep medicine: JCSM: Official publication of the American Academy of Sleep Medicine, 7*, S22.

NCSDR/NHTSA. (1998). Drowsy driving and automobile crashes. In E. P. o. D. F. a. Sleepiness (Ed.), *Drowsy driving and automobile crashes*. USA: National Center on Sleep Disorders Research, National Highway Traffic Safety Administration.

Philip, P., Sagaspe, P., Moore, N., Taillard, J., Charles, A., & Guilleminault, C. (2005). Fatigue, sleep restriction and driving performance. *Accident Analysis & Prevention, 37*, 473–478.

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.

Pinto, J., Cardoso, J., Lourenço, A., & Carreiras, C. (2017). Towards a continuous biometric system based on ECG signals acquired on the steering wheel. *Sensors, 17*, 2228.

Pudil, P. (1994). Floating search methods in feature selection. *Pattern Recognition Letters, 15*, 1119–1125.

Rahman, H., Ahmed, M. U., & Begum, S. (2016). Non-contact Heart Rate Monitoring Using Lab Color Space. *Studies in Health Technology and Informatics, 224*, 46–53.

Rahman, H., Barua, S., & Shahina, B. (2015). Intelligent driver monitoring based on physiological sensor signals: Application using camera. In *Proceedings of the IEEE eighteenth international conference on intelligent transportation systems* (pp. 2637–2642).

Reyner, L. A., & Horne, J. A. (1998). Falling asleep whilst driving: Are drivers aware of prior sleepiness? *International Journal of Legal Medicine, 111*, 120–123.

Reyner, L. A., Wells, S. J., Mortlock, V., & Horne, J. A. (2012). 'Post-lunch' sleepiness during prolonged, monotonous driving - effects of meal size. *Physiology & Behavior, 105*, 1088–1091.

Rohit, F., Kulathumani, V., Kavi, R., Elwarfalli, I., Kecojevic, V., & Nimbarte, A. (2017). Real-time drowsiness detection using wearable, lightweight brain sensing headbands. *IET Intelligent Transport Systems, 11*, 255–263.

Rüger, M., Gordijn, M. C. M., Beersma, D. G. M., de Vries, B., & Daan, S. (2006). Time-of-day-dependent effects of bright light exposure on human psychophysiology: Comparison of daytime and nighttime exposure. *American Journal of Physiology - Regulatory, Integrative and Comparative Physiology, 290*, R1413–R1420.

Sahayadhas, A., Sundaraj, K., & Murugappan, M. (2012). Detecting driver drowsiness based on sensors: A review. *Sensors, 12*, 16937–16953.

Sandberg, D., Åkerstedt, T., Anund, A., Kecklund, G., & Wahde, M. (2011). 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.

Simon, M., Schmidt, E. A., Kincses, W. E., Fritzsche, M., Bruns, A., Aufmuth, C., et al. (2011). EEG alpha spindle measures as indicators of driver fatigue under real traffic conditions. *Clinical Neurophysiology, 122*, 1168–1178.

Sparrow, A. R., LaJambe, C. M., & Van Dongen, H. P. A. (2018). Drowsiness measures for commercial motor vehicle operations. *Accident Analysis & Prevention*. doi:10.1016/j.aap.2018.04.020.

Vapnik, V. N. (1992). Principles of risk minimization for learning theory. *Advances in Neural Information Processing Systems, 4*, 831–838.

Vicente, J., Laguna, P., Bartra, A., & Bailón, R. (2016). Drowsiness detection using heart rate variability. *Medical & Biological Engineering & Computing, 54*, 927–937.

Wang, X., & Xu, C. (2016). Driver drowsiness detection based on non-intrusive metrics considering individual specifics. *Accident Analysis & Prevention, 95*, 350–357.

Watling, C. N., Armstrong, K. A., & Radun, I. (2015). Examining signs of driver sleepiness, usage of sleepiness countermeasures and the associations with sleepy driving behaviours and individual factors. *Accident Analysis & Prevention, 85*, 22–29.

Watson, I., & Marir, F. (1994). Case-based reasoning: A review. *The Knowledge Engineering Review, 9*, 327–354.

Welch, P. (1967). The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Transactions on Audio and Electroacoustics, 15*, 70–73.

Whitney, A. W. (1971). A direct method of nonparametric measurement selection. *IEEE Transactions on Computers, C-20*, 1100–1103.

Wierwille, W. W., Wreggit, S. S., Kirn, C. L., Ellsworth, L. A., & Fairbanks, R. J. (1994). Research on vehicle-based driver status/performance monitoring; development, validation, and refinement of algorithms for detection of driver drowsiness.

Yang, G., Lin, Y., & Bhattacharya, P. (2005). A driver fatigue recognition model using fusion of multiple features. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics: 2* (pp. 1777–1784). 1772.

Yang, Y., Kyrgyzov, O., Wiart, J., & Bloch, I. (2013). Subject-specific channel selector for classification of motor imagery electroencephalographic data. In *Proceedings of the thirty-eighth international conference on acoustics, speech, and signal processing* (pp. 1277–1280).

Yeo, M. V. M., Li, X., Shen, K., & Wilder-Smith, E. P. V. (2009). Can SVM be used for automatic EEG detection of drowsiness during car driving? *Safety Science, 47*, 115–124.

Zhang, C., Wang, H., & Fu, R. (2014). Automated detection of driver fatigue based on entropy and complexity measures. *IEEE Transactions on Intelligent Transportation Systems, 15*, 168–177.

Zhao, C., Zheng, C., Zhao, M., Tu, Y., & Liu, J. (2011). Multivariate autoregressive models and kernel learning algorithms for classifying driving mental fatigue based on electroencephalographic. *Expert Systems with Applications, 38*, 1859–1865.

Zheng, Y. L., Ding, X. R., Poon, C. C. Y., Lo, B. P. L., Zhang, H., Zhou, X. L., et al. (2014). Unobtrusive sensing and wearable devices for health informatics. *IEEE Transactions on Biomedical Engineering, 61*, 1538–1554.