Sensitivity and Specificity of the Driver Sleepiness Detection Methods using Physiological Signals: A Systematic Review

Christopher N. Watling[1,2], Md Mahmudul Hasan[1,2], Grégoire S. Larue[1,2]

[1] Queensland University of Technology (QUT), Centre for Accident Research and Road Safety - Queensland (CARRS-Q), Australia

[2] Queensland University of Technology (QUT), Institute of Health and Biomedical Innovation (IHBI), Australia

Corresponding author: christopher.watling@qut.edu.au
K Block, 130 Victoria Park Road
Kelvin Grove, QLD, 4059, Australia
Email: christopher.watling@qut.edu.au
Tel: +61731387747

Keywords: fatigue, drowsiness, driving, features, machine learning, ground truth, physiological sleepiness

Abstract

Driver sleepiness is a major contributor to road crashes. A system that monitors and warns the driver at a certain, critical level of arousal, could aid in reducing sleep-related crashes. To determine how driver sleepiness detection systems perform, a systematic review of the sensitivity and specificity outcomes was performed. In total, 21 studies were located that met inclusion criteria for the review. The range of sensitivity outcomes was between 39.0-98.8% and between 73.0-98.9% for specificity outcomes. There was considerable variation in the outcomes of the studies employing only one physiological measure (mono-signal approach), whereas, a poly-signal approach with multiple physiological signals resulted in more consistency with higher outcomes on both sensitivity and specificity metrics. Only six of the 21 studies had both sensitivity and specificity outcomes above 90.0%, which included mono- and poly-signal approaches. Moreover, increases in the number of features used in the sleepiness detection system did not result in higher sensitivity and specificity outcomes. Overall, there was considerable variability between the studies reviewed, including measures of ground truth, the features employed and the machine learning approach of the systems. A critical need for progressing any system is a revalidation of the system on a new sample of users. These aspects indicate considerable progress is needed with physiological-based driver sleepiness systems before they are at a sufficient standard to be deployed on-road.

1. Introduction

Driver sleepiness is a major issue for road safety (Australian Transport Council, 2011; Connor et al., 2002; Tefft, 2014). It has been noted that approximately 17.0% of fatal crashes in the USA (1999-2008) (Tefft, 2014) and 20.0-30.0% of road crash deaths and severe injuries in Australia (Australian Transport Council, 2011) are due to driver sleepiness. Thus, concerted efforts are necessary to mitigate its contribution to crash statistics.

Commonly, sleepiness is caused by insufficient sleep, although other factors such as time-on-task (Zeller et al., 2020), reduced stimulation from the environment (Larue et al., 2011) as well as environmental light levels (Ahlström et al., 2018), circadian-related factors or having a sleep disorder (Williamson et al., 2011), or use of medications (National Highway Traffic Safety Administration, 1998) can lead to reductions in alertness. Experiencing sleepiness is also a naturally occurring event governed by two intrinsic factors: a homeostatic component and a circadian component (Borbely, 1982). The homeostatic component is an increasing need to sleep that gradually develops from the time of awakening, whereas the circadian component promotes alertness in the morning with sleepiness gradually increasing from 14:00 and reaching a maximum sleep pressure during the hours of 02:00-06:00. A substantial proportion of drivers (59.0-77.0%) have reported feeling sleepy while driving (Armstrong et al., 2010; Vanlaar et al., 2008; Watling et al., 2015), indicating the behaviour is widespread. Moreover, certain driver types such as professional drivers, including bus drivers (Miller et al., 2020), heavy vehicle and taxi drivers (Meng et al., 2015), also commonly reported feeling sleepy while driving. Considered together, driver sleepiness is a prevalent and critical road safety issue. Thus, detecting and alerting a sleepy driver so they can implement a sleepiness countermeasure (i.e., nap or coffee consumption) would be beneficial for mitigating risk associated with sleepy driving.

Recent research on driver sleepiness detection has used data from three areas, including vehicle-based measures, behavioural measures, and physiological measures (Sahayadhas et al., 2012). Vehicle-based measures include wheel movement, pedal and accelerator movement, while behavioural-based measures comprise the detection of facial changes such as eye movements, yawning or changes in speech (Fan et al., 2007; Sahayadhas et al., 2012; Zhang & Zhang, 2010). There are several limitations of these two data sources as they are impacted by factors such as road markings, lighting conditions, climatic conditions, vehicle kinetic characteristics, facial characteristics, visual distraction and presence of a passenger (Choudhary

et al., 2017; Sahayadhas et al., 2012). Ongoing advances in hardware, as well as deep learning techniques, have led to improvements in behaviour-based measures. However, physiological measures are suggested as an effective measure for driver sleepiness detection (Doudou et al., 2019; Lal & Craig, 2001; Ramzan et al., 2019). Physiological signals including cortical activity via electroencephalography (EEG), various eye-related metrics from electrooculography (EOG) or infrared reflectance, and cardiac-related measures via electrocardiography (ECG) or pulse plethysmography and oximetry (Choudhary et al., 2016; Lal & Craig, 2001), have been found to reflect changes in arousal.

A number of approaches and methods have been used with the detection of sleepiness, including the mono-signal approach or the poly-signal approach being the use of one or several physiological signals, respectively. Moreover, a number of features can be extracted from the physiological signals; however, it is unclear which features might have more prominence with a well performing detection system. That is, while EEG-based measures of sleepiness such as alpha and theta power bands have been consistently related to increases in sleepiness, entropy-based measures (e.g., Min et al., 2017) and wavelet packet transformations (e.g., Chen et al., 2019) of the EEG signal also show promise with the assessment of sleepiness. Nevertheless, it is possible that the combination of different features might also improve detection outcomes, as sleepiness is a multifaceted state.

The most important aspect of a detection system is the sensitivity and specificity metrics to determine the overall performance of the system (Parikh et al., 2008). High sensitivity indicates that a system is highly capable of detecting the factor of interest, whereas high specificity signifies the ability of the system to identify the undesired factors more correctly. In the case of sleepiness detection systems, sensitivity is the estimate of the successful detection of sleepiness, and specificity is the estimate of the successful detection of alert state (Parikh et al., 2008). As the two metrics have implications for the efficacy of a system, it is important to consider these metrics with driver sleepiness detection systems.

Several driver sleepiness detection studies have some notable discrepancies with sensitivity and specificity of their proposed systems. For example, (Pritchett et al.'s 2011) detection system resulted in high sensitivity of 95.4% but a substantially lower specificity value of 75.8%. In this case, higher sensitivity and lower specificity characteristics can lead to the possibility of false alarms. On the other hand, Liang et al. (2019) achieved a low sensitivity of 39.0% and a

higher specificity of 98.0% while detecting driver sleepiness utilising the EEG signal. In this instance, there is a possible lower false alarm, but it is likely to misclassify sleepy states.

As sleepiness detection systems will be implemented as a part of the advanced driver-assistance systems, assessing the sensitivity and specificity of the detection models is a critical factor (Dawson et al., 2014). Previous work undertaken on this topic focused on describing driver sleepiness detection techniques with a limited examination of physiological signals (Sahayadhas et al., 2012) or the performance measures of any of the detection models (Saini & Saini, 2014), or included a high-level detailed review of detection technology but did not focus on classification, sensitivity or specificity (Dawson et al., 2014). Moreover, none of these studies followed the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA; Moher et al., 2009).

Understanding sensitivity and specificity outcomes of sleepiness detection systems provides a gauge of the success of these systems and analytical techniques. In addition, the utility of the classifiers and their relationship with sensitivity and specificity metrics can be assessed. Therefore, systematically reviewing empirical studies that have measured the sensitivity and specificity of their detection system is an important issue. The PRISMA criteria (Moher et al., 2009) guided the following review, including the development of the research questions, and consideration of participants, interventions, comparators, and outcomes. Thus, the following research questions were developed: i) what is the sensitivity and specificity of the driver sleepiness detection methods using physiological signals? and ii) does the number of features extracted from the physiological signals impact on the sensitivity and specificity metrics?

2. Methods

2.1. Data sources

A total of five databases (i.e., PubMed, Embase, Scopus, Web of Science and IEEE Xplore Digital Library) were searched in this systematic review. Moreover, other relevant literature identified from the initial search was manually screened. No starting date restriction was used given the recency of machine learning paradigms and their practical application. The end date was September 2020.

2.2. Study selection

Electronic databases were searched, and the title, abstracts and full-text documents were downloaded to EndNote for screening purposes. The screening and selection were performed

independently by two of the authors (CNW, MMH) and was validated by the third author (GSL). Disagreements on study inclusion was resolved via discussions between the three authors (CNW, MMH, and GSL).

2.3. Eligibility criteria

First, the terms for searching the databases, 'drowsiness', 'sleepiness' and 'fatigue', were utilised by using the 'OR' function. Second, for physiological signals, 'biosignal', 'EEG', 'EOG', 'ECG', 'EMG', 'skin conductance' and 'pulse oximetry' were used. Third, for the performance measure, 'sensitivity', 'specificity', and their synonyms (sensitivity: 'recall', 'true positive rate'; specificity: 'true negative rate') were used. After downloading all the articles with abstracts, further screening was done in the bibliographic software (EndNote), where studies were excluded other than the search results from the ANDing of 'driver', 'sleepiness', 'sensitivity' and 'specificity' (with their synonyms). This was performed to ensure only studies focusing on driver sleepiness were selected for review. The search strategy based on the PRISMA flow diagram is shown in Figure 1. All published and unpublished articles were considered, and there were no restrictions on the age or gender of the subjects. Moreover, there were no restrictions on language or location of data collection.

2.4. Data Extraction

The data extraction was performed by two authors (CNW and MMH), and was further validated by the third author (GSL). The information extracted from the selected publications included demographics, driving setting (simulator or on-road), data/measures used, extracted features, classification approach, sensitivity and specificity outcomes. A total of 69 articles were found, which specifically included the driver, sensitivity and specificity (and their synonyms). These articles were manually searched, and a total of 18 articles were excluded due to classifying sleep stages or stress level detection, being signal detection method papers, speech-, respiratory-, or vehicle-based detection methods. Articles that included at least one physiological measure and other signals such as respiration were still included in the review. Full-text articles of the remaining 51 studies were manually assessed for analysis. As 30 articles did not report the sensitivity, specificity or other vital aspects (e.g., ground truth, features extracted, classification method) of their proposed sleepiness detection systems, they were excluded as well, leaving 21 articles for review.

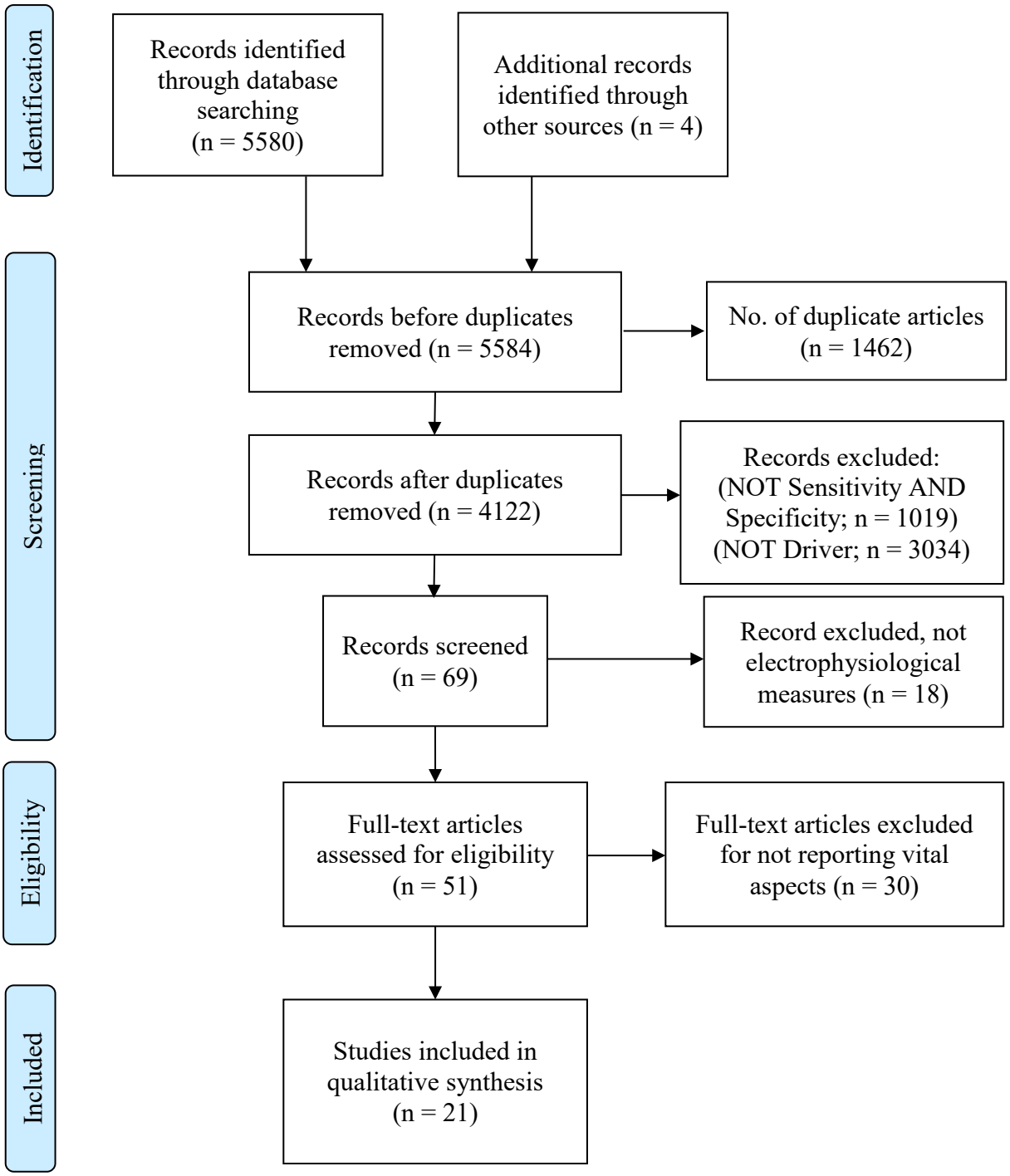

Figure 1: PRISMA-based flow diagram for the retrieval of studies on detection of driver sleepiness via physiological measures. Columns

3. Results

Table 1 displays the 21 studies included in the final study synthesis. Out of 21 studies, 17 studies used a mono-signal approach: 13 studies used EEG features (Chai et al., 2017; Chai et al., 2016; Chai et al., 2015; Chen et al., 2018a; Chen et al., 2018b; Chen et al., 2019; He et al., 2014; King et al., 2006; Ko et al., 2020; Liang et al., 2019; Mardi et al., 2011; Min et al., 2017;

Pritchett et al., 2011), two studies used ECG features (Persson et al., 2020; Vicente et al., 2016), one study used EOG features (Akerstedt et al., 2010) and one study used EMG features (Mahmoodi & Nahvi, 2019). A poly-signal approach was used by four studies, including the combination of EEG and EOG (Barua et al., 2019), EEG and ECG (Guo et al., 2016), EEG, EOG and ECG (Martensson et al., 2019), and EDA and PO (Bundele & Banerjee, 2010). It is worth noting that substantial variability is evident with the measures of ground truth employed (e.g., observer ratings of video, subjective sleepiness, performance indices), the classification models utilised, as well as the features extracted, even between studies that have used the same physiological signal (e.g., EEG: spectral power bands, wavelet transformations, entropies metrics).

Table 1: Details of the reviewed studies including demographics, study setting, data/measures used, extracted features, classification approach, sensitivity and specificity outcomes.

Study	Sample and demographics	Study Settings	Data/measures	Final model features	Classifiers	Sensitivity	Specificity
Ko et al. (2020)	N = 15 participants, aged: 22-28	Driving Simulator Driving route: highway driving (60 mins)	*Ground truth*: response times via steering wheel corrections *Input data:* physiological (EEG) and blink frequency *Classification time window:* 1 min	*4 features: EEG:* delta, theta, alpha and beta EEG power bands as well as blink frequency, duration, amplitude	*One classifier:* multiple regression model *Validation:* leave-one-trial-out cross-validation *Classes:* 2 (alert and fatigued)	58.0%	73.0%
Persson et al. (2020)	N = 86 sleep-deprived participants: Session 1: 18 (M:10, F:8); mean age: 41 Session 2: 24 (M:12; F:12); mean age: 35 Session 3: 44 (M:23; F:21); mean age: 44	On-road driving Session 1: Motorway (90 mins) Session 2: Motorway (135 mins) Session 3: Rural road (90 mins)	*Ground truth*: subjective sleepiness (KSS) *Input data:* physiological (ECG) *Classification time window:* 5 min	*5 features:* RMSSD, mean NN, LF_{abs}, SSD1, SSD2	*Four classifiers:* KNN, Gaussian kernel SVM, AdaBoost, random forest *Validation*: all classifiers were trained using 10-fold cross-validation *Data partitioning:* feature selection set (30 %), training set (50 %) and test set (20 %) using a holdout approach *Classes:* 3 (alert (KSS ≤ 5), somewhat sleepy (6 ≤ KSS ≤ 7), and severely sleepy (KSS ≥ 8)	57.9% random forest	79.3% random forest
Barua et al. (2019)	N = 30 male participants, age range: 18-25	Driving simulator Driving route: i) rural road-daylight; ii) rural road-night-time; iii) suburban road-daylight Driving session: 3 x 30-min driving	*Ground truth*: subjective sleepiness (KSS) *Input data:* physiological (EEG, EOG); PERCLOS, sleep-wake timing *Classification time window:* 5 min	*13 features:* EEG: δ-PSD, θ-PSD, α-PSD, β-PSD, (θ+α)/ β, α/ β, (θ+ α)/(α+ β), θ/β; EOG: blink duration; PERCLOS; sleep/wake predictor; driving condition: road type (rural or suburban road), lighting (daylight or night-time conditions)	*Four classifiers:* KNN, SVM, case-based reasoning, random forest *Validation:* 10-fold cross-validation and leave-one-out validation *Classes: 2* (alert, sleepy); a multiclass also trialled but resulted in poorer outcomes	85.0% case-based reasoning	81.0% case-based reasoning
Chen et al. (2019)	N = 16 male participants, mean age: 28.0	On-road driving Driving route: Driving session: 60-min driving	*Ground truth:* arbitrary classification of alert or fatigued based on first 3 min and last 3 min of driving respectively, confirmed by KSS values *Input data:* physiological (14 channel EEG) *Classification time window:* 3 min	*4 features:* Wavelet packet transform (WPT) of EEG data into δ-WPT, θ-WPT, α-WPT, β-WPT	*One classifier:* SVM *Validation:* 10-fold cross-validation *Classes:* 2 (alert or fatigued)	94.6%	94.3%

Study	Sample and demographics	Study Settings	Data/measures	Final model features	Classifiers	Sensitivity	Specificity
Martensson et al. (2019)	N = 86 sleep-deprived participants: Session 1: 18 (M:10, F:8); mean age: 41 Session 2: 24 (M:12; F:12); mean age: 35 Session 3: 44 (M:23; F:21); mean age: 44	On-road driving Session 1: Motorway (90 mins) Session 2: Motorway (135 mins) Session 3: Rural road (90 mins)	*Ground truth*: subjective sleepiness (KSS) *Input data:* physiological (EEG, EOG, ECG), sleep-wake timing *Classification time window:* 5 min	*10 features:* sleep/wake predictor; 5 EEG: θ/β (Cz-A2), θ/β (Oz-Pz), θ/(θ+α) (Fz-A1), α power (Cz-A2), α power (Fz-A1); EOG: mean lid closure speed, 90th percentile blink duration; ECG: RMSSD, Higuchi dimension	*Five classifiers:* random forest; AdaBoost, KNN, Linear SVM, Gaussian SVM *Validation*: 10-fold cross-validation *Data partitioning:* feature selection set (30%), training set (50%) and test set (20%) using a holdout approach *Classes:* 2 (severely sleepy (KSS≥8) or as sufficiently alert (KSS≤6))	86.5% random forest	95.7% random forest
Mahmoodi and Nahvi (2019)	N = 13 male participants, age range: 26-50	Driving simulator Driving route: 67 km closed-loop highway Driving speed: 80-100 km/h	*Ground truth*: observer rating of sleepiness *Input data:* physiological (EMG) *Classification time window:* 30 sec	*5 features:* EMG: range, variance, relative spectral power, kurtosis, and shape factor of EMG data	*Six classifiers:* KNN, regression tree, binary SVM, naïve bayes model, ensemble of learners for regression, SVM regression model *Validation*: 20-fold cross-validation *Data partitioning:* 70% for the training set, 15% for validation, and 15% for testing test *Classes:* 2 (alert or drowsy)	77.0% KNN	92.0% KNN
Liang et al. (2019)	N = 16 participants (M:7; F:9) night shift workers, mean age: 48.7	On-road driving Driving route: two-lane 0.8 km closed-loop Driving session: 2 x 2-h driving	*Ground truth*: model 1: microsleeps (EEG) model 2: lane crossing events *Input data:* Vehicle data and ocular data (Optalert™) *Classification time window:* *Model 1:* 1 min Model 2: 10 min	*10 features:* vehicle data: SD of land position, SD of steering wheel position, mean steering wheel error; ocular data: mean and SD of amplitude/ velocity ratio (AVR), mean and SD of negative AVR, % Eye Closure (PERCLOS), John's drowsiness score (JDS); individual driver factor	*One classifier:* logistic regression models *Validation:* 75% training and 25% testing via a random spilt *Classes:* 2 (non-drowsiness or drowsiness)	Model 1: 39.0% Model 2: 36.0%	Model 1: 98.0% Model 2: 98.0%
Chen et al. (2018a)	N = 12 male participants, mean age: 27.33	Driving simulator Driving route: monotonous driving scenario Driving session: 60-min driving	*Ground truth:* arbitrary classification of alert or fatigued based on first 3 min and last 3 min of driving respectively, confirmed by KSS values *Input data:* physiological (30 channel EEG) *Classification time window:* 3 min	*4 features:* EEG: delta, theta, alpha and beta EEG power bands	*One classifier:* novel fusion feature (FBN-PSD-FF) and Extreme learning machine (ELM) *Validation:* leave-one-participant-out-cross-validation *Classes:* 2 (alert or fatigued)	95.7%	94.3%

Study	Sample and demographics	Study Settings	Data/measures	Final model features	Classifiers	Sensitivity	Specificity
Chen et al. (2018b)	N = 15 participants, mean age: 26.2	Driving simulator Task 1: crowded four-lane road Task 2: monotonous driving two-lane road	*Ground truth:* arbitrary classification of alert or fatigued based on first 2 min and last 2 min of driving respectively, confirmed by KSS values *Input data:* physiological (30 channel EEG) *Classification time window:* 2 min	*4 features:* EEG: delta, theta, alpha and beta EEG power bands	*Four classifiers:* Gaussian SVM, KNN, logistic regression, decision trees *Validation:* 10-fold cross-validation *Classes:* 2 (alert and drowsy state)	98.8% KNN	98.9% KNN
Chai et al. (2017)	N = 43 participants, age range: 18-55	Divided Attention Steering Simulator (DASS) Driving session: up to 2 hrs	*Ground truth:* arbitrary classification of alert-fatigued based on first and last 5 min of driving respectively, 20 sec of data selected classification from each 5 min period *Input data:* physiological: (32 channel EEG), behavioural: reaction time *Classification time window:* 2 sec	*128 features:* 32 EEG channels x δ-PSD, θ-PSD, α-PSD, β-PSD with autoregressive modelling	*Four classifiers:* ANN, Bayesian neural network, deep belief network, sparse-deep belief networks *Validation:* hold-out cross-validation and k-fold cross-validation. *Data partitioning:* 33% training, validation 33%, and testing 34% *Classes:* 2 (alert or fatigued)	93.9% sparse-deep belief networks	92.3% sparse-deep belief networks
Min et al. (2017)	N = 12 sleep deprived participants (M: 12, F:0), age range: 19-24	Driving simulator Driving session: 1-2 hours of highway driving - low traffic density	*Ground truth:* arbitrary classification of alert or fatigued based on first 5 min and last 5 min of driving respectively, confirmed by Chalder Fatigue Scale, Li's Subjective Fatigue Scale, visual signs of sleepiness, crashes and lane deviations *Input data:* physiological (30 channel EEG) *Classification time window:* 5 min	*4 features:* EEG: spectral entropy, approximate entropy, sample entropy and fuzzy entropy	*Four classifiers:* radial basis functions SVM, back-propagation neural network (BPNN), random forest, KNN *Validation:* leave-one-out cross-validation approach *Data partitioning:* 50% training and 50% testing datasets (random allocation) *Classes:* 2 (normal and a fatigued state)	98.3% BPNN	98.2% BPNN

Study	Sample and demographics	Study Settings	Data/measures	Final model features	Classifiers	Sensitivity	Specificity
Chai et al. (2016)	N = 43 participants, age range: 18-55	Divided Attention Steering Simulator (DASS) Driving session: up to 2 hrs	*Ground truth:* arbitrary classification of alert-fatigued based on first and last 5 min of driving respectively, 20 sec of data selected classification from each 5 min period *Input data:* physiological: (32 channel EEG), behavioural: reaction time *Classification time window:* 2 sec	*128 features:* 32 EEG channels x δ-PSD, θ-PSD, α-PSD, β-PSD with autoregressive modelling	*One classifier:* Bayesian neural network *Validation:* validation set not used with inclusion of hyperparameters in the cost function of the Bayesian regularisation algorithm *Data partitioning:* 50% training and 25% testing sets via random split *Classes:* 2 (alert or fatigued)	89.7%	86.8%
Vicente et al. (2016)	N = 30 participants, some sleep deprived, (M: 17, F:13), age range: 25-60	Driving simulator and on-road driving Session 1 (Simulator): n = 9, 120 mins Session 2 (Simulator): n = 11, 100 mins Session 3 (on-road): n = 8, 8h driving, stop every 2 hrs for 10 min	*Ground truth:* model 1: awake or drowsy via observer ratings model 2: sleep deprived or not-sleep deprived *Input data:* physiological (ECG) *Classification time window:* 1 min (both models)	7 *features:* ECG: heart rate, absolute and normalised LF and HF powers, LF/HF ratio, respiratory frequency and percentage of total power (ξ)	*One classifier:* linear discriminant analysis *Validation:* leave-one-participant-out-cross-validation *Classes:* Model 1, 2 (awake or sleepy) and Model 2, (sleep deprived or not-sleep deprived)	Model 1: 59.0% Model 2: 62.0%	Model 1: 98.0% Model 2: 88.0%
Guo et al. (2016)	N = 20 participants, (M:12, F:8), age range: 24-51	Driving simulator Driving session: 4-6 hrs freeway driving	*Ground truth:* subjective sleepiness, via Stanford Sleepiness Scale (SSS) *Input data:* physiological: (EEG, ECG), behavioural: reaction time *Classification time window:* 3 min	*9 features:* EEG: α-PSD, β-PSD, δ-PSD, EEG-PSD (1-30 Hz), α/β, (α+θ)/ β, α-PSD/β-PSD; ECG: heart rate, RR_{SD} (Standard deviation (SD) of the RR interval)	*One classifier:* genetic algorithm-based SVM *Validation:* 10-fold Cross-Validation *Data partitioning:* 33% training, validation 33%, and testing 34% *Classes:* 2 (alert or sleepy)	87.5%	85.5%
Chai et al. (2015)	N = 43 participants, age range: 18-55	Divided Attention Steering Simulator (DASS) Driving session: up to 2 hrs	*Ground truth:* arbitrary classification of alert-fatigued based on first and last 5 min of driving respectively *Input data:* physiological: (32 channel EEG), behavioural: reaction time *Classification time window:* 5 min	*128 features:* 32 EEG channels x δ-PSD, θ-PSD, α-PSD, β-PSD	*One classifier:* Fuzzy swarm based-artificial neural network (ANN) *Data partitioning:* 33% training, validation 33%, and testing 34% *Classes:* 2 (alert or fatigued)	78.2%	79.6%

Study	Sample and demographics	Study Settings	Data/measures	Final model features	Classifiers	Sensitivity	Specificity
He et al. (2014)	N = 30 participants, age range: 18-43	Driving simulator Driving session: 1 hr	*Ground truth:* observer ratings of fatigued state *Input data:* physiological: (MindWave EEG headset) *Classification time window:* not reported	*2 features:* EEG: attention and meditation metrics based on propriety algorithms (NeuroSky, Inc, California)	*One classifier:* KNN *Data separation:* 12 participants for training and 18 participants for testing *Classes:* 2 (alert or fatigued)	68.3%	90.4%
Mardi et al. (2011)	N = 10 participants (M: 7, F: 3), mean age: 27.7	Driving simulator Driving session: 45 min	*Ground truth:* driving performance and observer rating *Input data: physiological: (24 channel EEG)* *Classification time window:* 2 sec	*3 features: Higuchi's Fractal Dimension, Petrosian's Fractal Dimension, Logarithm of Energy of Signal*	*One classifier:* ANN *Data separation:* 80% training and testing 20% *Classes:* 2 (alert or drowsy)	83.8%	84.9%
Pritchett et al. (2011)	N = 45 participants, age range: 20-60	Driving simulator Driving session: 2.5 h (starting from 2:30 pm)	*Ground truth:* observer ratings of sleepiness state *Input data:* Physiological: EEG, body movement data via piezoelectric sensors *Classification time window:* 1 min	*6 features:* EEG: alpha burst duration, current alpha wave count, minimum alpha wave count, wave duration variance, slope smoothness measurement; Body movement: average peak-to-peak body movement	*Two classifiers:* single and hybrid source algorithm. *Classes:* 2 (alert or sleepy)	95.4%	75.8%
Bundele and Banerjee (2010)	N = 10 participants, professional drivers, age range: 25-55	On-road driving Driving session: not reported	*Ground truth:* arbitrary classification of alert or fatigued based on pre-driving and post-driving values respectively *Input data:* Physiological: EDA, PO *Classification time window:* not reported	*16 features:* skin conductance and pulse oximetry: mean of signal, SD of signal, frame energy, maximum frequency, SD of frequency spectrum, mean of frequency spectrum, gradient, slope	*One classifier:* Multilayer perceptron neural network (MLPNN) *Data partitioning:* 50% training, validation 25%, and testing 25% *Classes:* 2 (alert or fatigued)	94.1% MLPNN	97.3% MLPNN
Åkerstedt et al. (2010)	N = 13 participants (50% male), mean age: 37.9, range: 24-57	Driving simulator Driving session: six 1 h sessions over a 24 h period, when rested and partially sleep deprived	*Ground truth:* model 1: subjective sleepiness (KSS) model 2: driving performance *Input data:* physiological (EOG), driving performance *Classification time window: 5* min (both models)	*3 features:* EOG: blink duration, blink amplitude/peak closing velocity, driving performance: SDLP	*One classifier:* linear discriminant analysis *Classes:* Model 1, 2 (awake or severe sleepiness (KSS ≥ 8)); Model 2, 2 (driving with no line crossing or line crossing – 2 wheels crossing outer edge of centre line)	Model 1: 73.0% Model 2: 23.0%	Model 1: 71.0% Model 2: 99.0%

Study	Sample and demographics	Study Settings	Data/measures	Final model features	Classifiers	Sensitivity	Specificity
King et al. (2006)	N = 55 participants n = 20 professional drivers, mean age range: 44 (SD = 11) n = 20 non-professional drivers, mean age range: 34 (SD = 21)	Driving simulator Driving session: drive until judged fatigued	*Ground truth:* observer ratings of drivers fatigued state *Input data:* Physiological: 19 channel EEG *Classification time window:* 1 min	*4 features:* delta, theta, alpha and beta EEG power bands	*One classifier:* ANN *Data partitioning:* 60% training, validation 15%, and testing 15% *Classes:* 2 (alert or fatigued)	80.5% (pro drivers) 84.0% (non-pro drivers)	82.4% (pro drivers) 82.1% (non-pro drivers)

Note: EEG, electroencephalography; ECG, electrography; EMG, electromyography; EDA, electrodermal activity; PO, pulse oximetry; PSD, power spectrum density; WPT, Wavelet packet transform; PERCLOS, percentage of eyelid closure; JDS, John's drowsiness score; KSS, Karolinska Sleepiness Scale; SSS. Stanford Sleepiness Scale; RMSSD, root mean square of successive RR interval differences; Higuchi dimension, a measure of irregularity; LF, low frequency; HF, high frequency; RR_{SD}, Standard deviation of the RR interval; NN, interval; SD1, Poincaré plot standard deviation perpendicular the line of identity; SD2, Poincaré plot standard deviation along the line of identity; SVM, support vector machine; KNN, k-nearest neighbours; ANN, artificial neural networks; BPNN, back-propagation neural network; DGM, hybrid deep generic model; MLPNN, multilayer perceptron neural network; DASS, Divided Attention Steering Simulator; SDLP, Standard deviation of lateral position.
The impaired state (sleepy/fatigued/drowsy) was the predicted outcome thus, sensitivity represents the correct detection of the impaired state across all studies.

3.1 Sensitivity and specificity outcomes

The sensitivity and specificity outcomes from the different studies highlight considerable variability across the studies. Figure 2 shows the sensitivity and specificity outcomes for each study which is arranged by the average of sensitivity and specificity as well as by mono-signal versus poly-signal approach. The use of a mono-signal approach has not been a particularly reliable approach to detect sleepiness. Studies using a mono-signal approach either have obtained a higher sensitivity with a lower specificity (i.e., Pritchett et al., 2011) or vice versa (i.e., Liang et al., 2019; Persson et al., 2020; Vicente et al., 2016). There are some exceptions where high, or a moderate level of both sensitivity and specificity (range = 82.1-98.9%) were achieved in the study (i.e., Chai et al., 2017; Chai et al., 2016; Chen et al., 2018a; Chen et al., 2018b; Chen et al., 2019; King et al., 2006; Mardi et al., 2011; Min et al., 2017) despite using a single physiological signal. Whereas, all the poly-signal approaches consistently achieved moderate to high sensitivity and specificity compared to the single physiological signal-based systems (Barua et al., 2019; Bundele & Banerjee, 2010; Guo et al., 2016; Martensson et al., 2019). Thus, the use of a poly-signal approach consistently provides both higher sensitivity and specificity outcomes and appears as a useful approach for improving sleepiness detection systems.

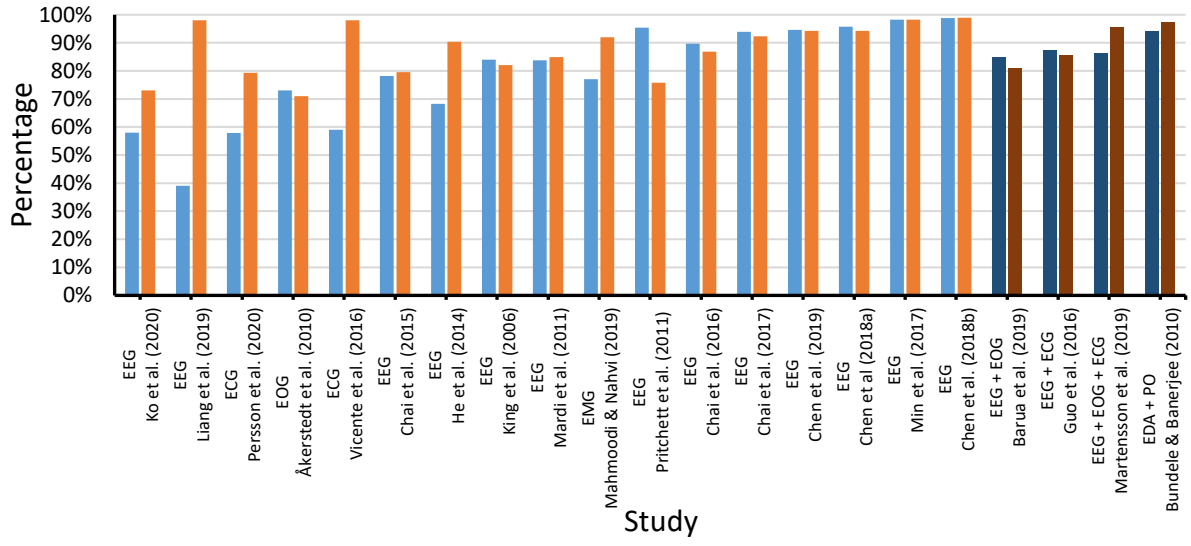

Figure 2: Sensitivity (blue bars) and specificity (orange bars) values of mono- (light colours) and poly-signal (darker bars) based sleepiness detection approaches. EEG, electroencephalography; ECG, electrography; EMG, electromyography; EDA, electrodermal activity; PO, pulse oximetry. Studies have been arranged by an average of sensitivity and specificity as well as by mono-signal versus poly-signal approach.

3.2 Extracted features

Figure 3 displays the relationship between sensitivity and specificity metrics plotted against the number of features extracted. The number of different features range from 4 to 13. It must be noted that using a greater number of features does not necessarily increase the sensitivity and specificity of the driver sleepiness detection system. For instance, Barua et al. (2019), using a mono-signal approach with 13 features, achieved an 85.0% sensitivity and 81.0% specificity, whereas, Chen et al. (2018b), using a mono-signal approach, used four features and obtained 98.8% sensitivity and 98.9% specificity. Variations of sensitivity and specificity also occur when the same number of features are used. Chen et al. (2018a) and King et al. (2006) both used a mono-signal approach with four features, yet obtained vastly different indices of sensitivity (95.7% versus 84.0%) and specificity (94.3% vs 82.1%), respectively. Considered together, there is no clear relationship between the number of features and the resultant sensitivity and specificity outcomes.

The use of features that have more relevance for variations in sleepiness seemingly provides better outcomes (e.g., EEG features; Chen et al., 2018b; King et al., 2006; Min et al., 2017, see Table 1) and could be an important consideration for improving outcomes with any detection model. However, there are some exceptions with using features that are more relevant with sleepiness. Pritchett et al. (2011) and Chai et al. (2015) reported using EEG defined sleepiness features in their studies' methods; however, the sensitivity and specificity outcomes (both < 80%) were not that impressive. Further, Bundele and Banerjee (2010) used features from electrodermal activity and pulse oximetry, and obtained sensitivity and specificity outcomes of 94.1% and 97.3% respectively. Suggesting that alternative measures of sleepiness may also have some utility for the detection of sleepiness.

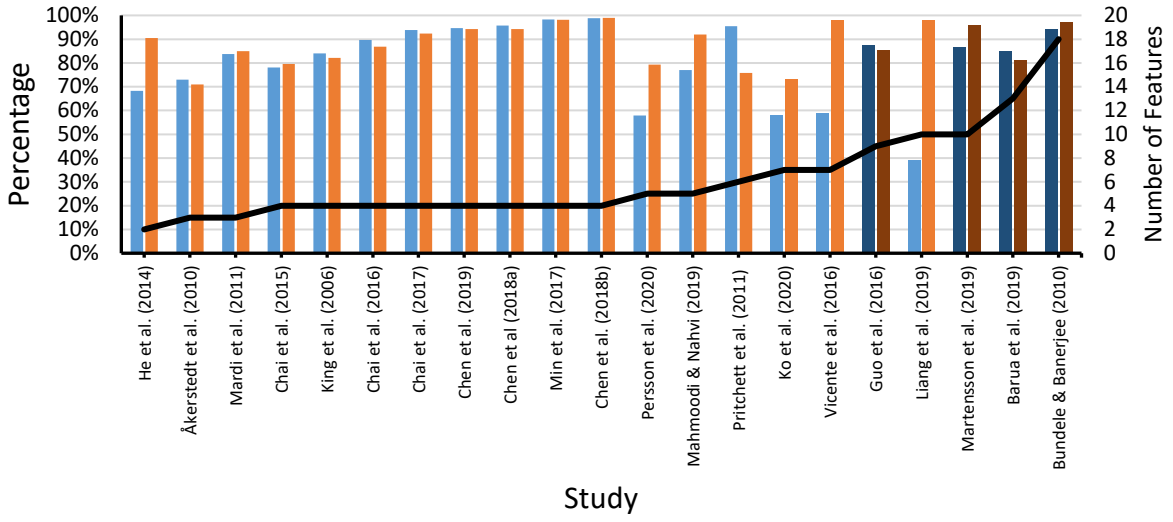

Figure 3: Sensitivity (blue bars) and specificity (orange bars) values of mono-signal (light colours) and poly-signal approaches (dark colours) of the studies included in the review with the number of features plotted also plotted (black line). Data order by the number of features.

4. Discussion

The current study sought to evaluate the sensitivity and specificity outcomes for studies examining the detection of driver sleepiness. Overall, the outcomes show substantial variability across the studies included in the review; however, some trends were apparent. For instance, sensitivity and specificity metrics were consistently higher when poly-signal approaches were used with less variability between sensitivity and specificity outcomes. There was also no consistency for sensitivity and specificity outcomes when consideration was given to the number of features used in the analysis.

4.1. Sensitivity and Specificity

The first aim of review was to determine the sensitivity and specificity of the driver sleepiness detection methods using physiological signals. The review highlights variations across the types of physiological measures used, as well as variations in the magnitude of sensitivity and specificity outcomes. Overall, a poly-signal approach (using multiple physiological measures) consistently provides higher values of sensitivity and specificity (i.e., Bundele & Banerjee, 2010; Guo et al., 2016; Martensson et al., 2019). This outcome is not surprising, especially when one considers the varied and multifaceted nature of sleepiness, as well as the diverse way in which sleepiness can present between individuals (e.g., Ingre et al., 2006; Van Dongen et al., 2012). The general trend was for poly-signal approaches to have higher sensitivity and specificity values with a smaller range between the two metrics. However, additional physiological signals does not guarantee higher sensitivity and specificity values. Such that, the studies of Barua et al. (2019) and Guo et al. (2016) had moderate sensitivity and specificity outcomes in the range of 81.0-87.5%.

The use of more signals does represent some theoretical and practical issues. Theoretical issues with the use of multiple sensors relate to individual differences regarding sleepiness impairment (e.g., Ingre et al., 2006; Van Dongen et al., 2012), but also differences between individuals on how this is reflected in physiological data (Schleicher et al., 2008). A concern related to individual differences is that of training of the driver sleepiness detection system and how best this can be achieved between individuals. Certainly, the response to sleepiness is suggested to be a stable, trait-like feature (Leproult et al., 2003; Van Dongen et al., 2004) and thus, once a system is tuned to a particular individual, could theoretically function without an

issue. Some initial progress with adapting a pre-trained artificial neural network for use on a new driver has provided positive results (i.e., de Naurois et al., 2018). Moreover, using baseline-corrected features during the training phase, as well as leave-one-out cross-validation, will likely improve the predictive utility across individuals. However, further work is seemingly needed to explore how best to account for interindividual differences in sleepiness.

Practical issues relate to transferring laboratory and field-based research outcomes to on-road applications and user acceptance. First, field-based research is likely to encounter several confounding factors such as environmental (i.e., sun glare and changes to ocular characteristics) and/or endogenous (i.e., a relaxed driver and changes to arousal levels) factors that might result in increased false positives or negatives. It has been noted that both subjective (i.e., KSS) and physiological (i.e., blink durations, Karolinska Drowsiness Scores) arousal levels are lower in driving simulator settings when compared to the on-road settings (Fors et al., 2018; Hallvig et al., 2013) which could present issues when transferring simulator-based detection models to real road environments. All of these issues will need meticulous and systematic assessment during testing and development. Moreover, laboratory and field-based studies allow for careful setup of physiological sensors and meticulous post-processing of physiological data, given the sensitive nature of physiological data and, consequently, signal artefact (Gratton, 2007). However, the use of multiple sensors is a major issue to be considered in the real-time implementation, especially for computational load and setup time (Balandong et al., 2018; Belakhdar et al., 2016) as well as user-friendliness, intrusiveness and ergonomics (Sahayadhas et al., 2012). In addition, multiple sensors will naturally increase the cost of the system (Patel et al., 2012), while reliable wearable technology and non-obtrusive sensors could alleviate or mitigate issues with user-friendliness, intrusiveness and ergonomics for the user. Nevertheless, research should focus on achieving an optimal balance of these noted factors, which will influence user acceptance.

4.2. Extracted Features

The current review also sought to examine the relationship between the number of features extracted on sensitivity and specificity metrics. Overall, no clear relationship between the number of features and the resultant sensitivity and specificity outcomes could be determined. The use of features that have more relevance to sleepiness seems to be a key consideration. Specifically, EEG, which is a measure of cortical arousal, has been a favoured physiological signal, as increases in theta and alpha power is consistently associated with increased sleepiness (Dawson et al., 2014).

A number of studies have used EEG data and extracted various features, including spectral analysis, on different power bands, frequency and wavelet transforms, nonlinear methods, and entropies (e.g., He et al., 2014; Khushaba et al., 2013; Min et al., 2017). Some features seem to be successful in obtaining higher sensitivity and specificity outcomes. For example, the fusion of the multiple entropies, that being spectral entropy, approximate entropy, sample entropy and fuzzy entropy extracted from EEG in the study of Min et al. (2017) are more successful than the EEG-α, β, and θ band powers extracted by Chai et al. (2015). Potential reasons for the higher outcomes could be the fusion of all the entropy-based features and selection of significant channel regions using their proposed channel selection method (Min et al., 2017). Moreover, the authors emphasised that the natural variability of the awaking EEG over time (i.e., entropy) naturally suits entropy metrics and facilitates higher sensitivity and specificity outcomes (Min et al., 2017).

4.3. Issues of consistency

The reviewed studies included a number of inconsistent findings that relate to the types of classifiers, the features extracted, the measure used as ground truth, and the classification time window. Several classifiers provided the higher estimates of sensitivity and specificity. In particular, random forest (Martensson et al., 2019; Persson et al., 2020), KNN (Chen et al., 2018b), SVM (Guo et al., 2016), and various neural networks classifiers (Bundele & Banerjee, 2010; Min et al., 2017) all provided the highest estimates of sensitivity and specificity. This does signify that these various learning techniques all have some utility with classification of sleepiness, however, which techniques holds the most promise is still unclear. There was also substantial variability with the extracted features employed and, along with the different classifiers, it is also difficult to consolidate which features are more relevant. For instance, considering the studies from Chen et al. (2018b) and Min et al. (2017), both studies achieved very high outcomes on sensitivity and specificity (range = 98.2-98.9) using EEG data but both extracted substantially different features (traditional power bands verses entropy features, respectively) and also used different classifiers (KNN and back-propagation neural network). Martensson et al. (2019)'s study included the traditional EEG power bands and entropy features; however, the entropy features did not make the cut off marks for inclusion in that study's final model.

Given the discrepancy noted above, it is possible that different design aspects could be an influential factor. Examining Table 1, several studies (i.e., Bundele & Banerjee, 2010; Chai et al., 2017; Chai et al., 2016; Chai et al., 2015; Chen et al., 2018a; Chen et al., 2018b; Chen et

al., 2019; Min et al., 2017) used an arbitrary classification approach with determining ground truth and used the first and last sections of the driving task when sleepiness would be at its lowest and highest. Aside from results found in Chai et al.'s (2015) study, the comparison of limited time windows of an alert and fatigued state resulted in very high sensitivity and specificity outcomes (between 94.1-98.9%). Although these outcomes are encouraging, the utility of such a design is extremely limited, as moment-to-moment monitoring of the driver's state should be the goal for any detection system.

The studies performed by Guo et al. (2016) and Martensson et al. (2019) employed classification time windows of 3 and 5 minutes respectively, over the entire driving session, and provided a more useful monitoring strategy of the drivers' state – sensitivity and specificity outcomes ranged between 84.5-95.7%. Improvements in sensitivity and specificity outcomes are, of course, required; however, the temporal resolution of the time windows does allow for a system to 'warn' the driver in a timely manner. Research on commercially available monitoring devices suggests that reliable outcomes were only obtained when a temporal resolution of greater than 30 minutes was used (e.g., Golz et al., 2010). It is possible that this could be problematic in certain driving situations (but not all).

The training and testing of the models also had some considerable differences across studies that likely influenced the overall outcomes. Common validation techniques noted were the k-fold cross-validation, hold-out validation, and leave-one-out (trial or participant) validation. Hold-out validation is known to overfit the data; however, the review noted a wide range of sensitivity and specificity values (36.0-98.3%) using this technique. Whereas, the k-fold cross-validation can limit issues with overfitting, yet, sensitivity and specificity outcomes ranged between 57.9-98.9% with similar ranges with the leave-one-out validation (58.0-98.3%). As such, variations in sensitivity and specificity outcomes with the different validation techniques could be due to the different validation techniques and/or linked to the issues previously noted above.

Perhaps the most salient issue relating to inconsistencies is the different measures used as ground truth. Measures of ground truth include reaction times and vehicle performance (i.e., Ko et al., 2020), and arbitrary classification of alert or fatigue based on pre-driving data and at the end of an extended driving period (e.g., Bundele & Banerjee, 2010; Chai et al., 2015; Chen et al., 2018a; Chen et al., 2019; Phyo Phyo et al., 2016). The majority of studies reviewed have employed observer ratings as the ground truth (see Table 1). However, issues associated with

observer ratings include low inter-rater agreements and poor correspondence with subjective sleepiness measures (Ahlstrom et al., 2015), such as the well-validated and widely used Karolinska Sleepiness Scale (Åkerstedt et al., 2014), which is also a commonly used measure of ground truth. Performance-based measures of ground truth could also be problematic, as motivational factors including exerting more effort to the task can, for a limited amount of time, lessen performance decrements associated with sleepiness (Boksem et al., 2006; Watling, 2016).

Subjective sleepiness measures, however, seem to hold some promise as a measure of ground truth. Martensson et al. (2019) employed a two-class system based on the participants subjective sleepiness scores via the 9-point Karolinska Sleepiness Scale (KSS; Åkerstedt & Gillberg, 1990). In this study, being sleepy was determined when KSS scores were ≥ 8, and being sufficiently alert was determined when KSS scores were ≤ 6. This procedure allowed for clear separation of the two classification states. Several studies show KSS values of 8 or more clearly demonstrate sleepiness-related impairment (Ingre et al., 2006; Watling et al., 2016) and, importantly, are predictive of a greater likelihood of on-road sleep-related crashes (Åkerstedt et al., 2008). Though, it should be noted that KSS ratings, like any subjective/self-report data, are subject to bias and can be influenced by the context (monotonous settings, light levels, social interaction) when ratings are obtained (Åkerstedt et al., 2014). Lastly, a driver's acceptance of a detection system could also be highly influenced by what the driver subjectively feels, which could also influence actions such as choosing to stop driving when too sleepy.

Overall, there are several inconsistencies across a range of study variables, including the driving environment, ground truth measures as well as time windows, and the physiological signals. As such, the results from different studies are problematic to compare with so many and varied differences, and this makes for a lack of generalisation when specific aspects (e.g., classifier or extracted features) are to be improved.

4.4 Limitations and Future Considerations

The current review is not without limitation and, as such, the findings need to be interpreted with the limitations in mind. A study limitation was the choice of key search terms, as other terms such as 'confusion matrix' could have been included in the review. As noted in the results, numerous data sources for the classifier and inputs have been used, numerous analytical techniques are used with feature extraction and classification, and all of these are not without

their limitations. Given the varied metrics and analytical techniques, comparisons of each were not always feasible. Future considerations necessitate some consistency with the measures used and the sources of ground truth, such as the well-validated KSS. Lastly, no study was found that has applied a detection system with a high level of sensitivity or specificity to a completely new sample for revalidation – this should be common procedure if detection systems are to be successful. A number of considerations should be examined with classification systems and the measure of ground truth.

4.5 Conclusion

In summary, sleep-related crashes account for a substantial proportion of road crash incidents. Driver sleepiness detection systems have the potential to warn drivers should their level of sleepiness increase, and can thus reduce the risk associated with sleep-related crashes. Detection systems that use a poly-signal approach seemingly have more utility for producing higher sensitivity and specificity values. The number of extracted features is of little consequence with sensitivity and specificity values; however, what is clear is the need to use features more relevant for sleepiness. As no individual is immune to the effects of sleepiness, developing systems that can aid drivers to make safer sleepy driving choices are vital.

Acknowledgements

The authors would like to sincerely thank Sonali Nandavar for her diligent proofreading of this article.

5. References

Ahlström, C., Anund, A., Fors, C., & Åkerstedt, T. (2018). The effect of daylight versus darkness on driver sleepiness: a driving simulator study. *27*(3), e12642.

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*(4), 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*(3), 240-252.

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*(4), 1480-1485.

Åkerstedt, T., & Gillberg, M. (1990). Subjective and objective sleepiness in the active individual. *International Journal of Neuroscience, 52*(1-2), 29-37.

Akerstedt, 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. *J Sleep Res, 19*(2), 298-309.

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*(2), 298-309.

Armstrong, K. A., Obst, P., Banks, T., & Smith, S. S. (2010). Managing driver fatigue: Education or motivation? *Road & Transport Research, 19*(3), 14-20.

Australian Transport Council. (2011). *National Road Safety Strategy 2011 – 2020*. T. Department of Infrastructure, Cities and Regional Development,.

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. (2019). Automatic driver sleepiness detection using EEG, EOG and contextual information. *Expert Systems With Applications, 115*, 121-135.

Belakhdar, I., Kaaniche, W., Djmel, R., & Ouni, B. (2016). A comparison between ANN and SVM classifier for drowsiness detection based on single EEG channel. 2016 2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP),

Boksem, M. A. S., Meijman, T. F., & Lorist, M. M. (2006). Mental fatigue, motivation and action monitoring. *Biological Psychology, 72*(2), 123-132.

Borbely, A. A. (1982). A two process model of sleep regulation. *Hum Neurobiol, 1*(3), 195-204.

Bundele, M. M., & Banerjee, R. (2010). ROC analysis of a fatigue classifier for vehicular drivers. 2010 5th IEEE International Conference Intelligent Systems,

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 [Original Research]. *11*(103).

Chai, R., Naik, G. R., Nguyen, T. N., Ling, S. H., Tran, Y., Craig, A., & Nguyen, H. T. (2016). Driver fatigue classification with independent component by entropy rate bound minimization analysis in an EEG-based system. *IEEE Journal of Biomedical and Health Informatics, 21*(3), 715-724.

Chai, R., Naik, G. R., Tran, Y., Ling, S. H., Craig, A., & Nguyen, H. T. (2015). Classification of driver fatigue in an electroencephalography-based countermeasure system with source separation module. 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC),

Chen, J., Wang, H., & Hua, C. (2018a). Electroencephalography based fatigue detection using a novel feature fusion and extreme learning machine. *Cognitive Systems Research, 52*, 715-728.

Chen, J., Wang, H., & Hua, C. C. (2018b). Assessment of driver drowsiness using electroencephalogram signals based on multiple functional brain networks. *International Journal of Psychophysiology, 133*, 120-130.

Chen, J., Wang, H., Wang, Q., & Hua, C. (2019). Exploring the fatigue affecting electroencephalography based functional brain networks during real driving in young males. *Neuropsychologia, 129*, 200-211.

Choudhary, P., Sharma, R., Singh, G., Das, S. J. I. R. J. o. E., & Technology. (2016). A survey paper on drowsiness detection & alarm system for drivers. *International Research Journal of Engineering, 3*(12), 1433-1437.

Choudhary, P., Velaga, N. R. J. T. r. p. F. t. p., & behaviour. (2017). Analysis of vehicle-based lateral performance measures during distracted driving due to phone use. *Transportation Research Part F: Traffic Psychology, 44*, 120-133.

Connor, J., Norton, R., Ameratunga, S., Robinson, E., Civil, I., Dunn, R., Bailey, J., & Jackson, R. J. B. (2002). Driver sleepiness and risk of serious injury to car occupants: population based case control study. *BMJ, 324*(7346), 1125.

Dawson, D., Searle, A. K., & Paterson, J. L. (2014). Look before you (s)leep: Evaluating the use of fatigue detection technologies within a fatigue risk management system for the road transport industry. *Sleep Medicine Reviews, 18*(2), 141-152.

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.

Doudou, M., Bouabdallah, A., & Berge-Cherfaoui, V. (2019). Driver Drowsiness Measurement Technologies: Current Research, Market Solutions, and Challenges. *International Journal of Intelligent Transportation Systems Research*, 1-23.

Fan, X., Yin, B.-C., & Sun, Y.-F. (2007). Yawning detection for monitoring driver fatigue. 2007 International Conference on Machine Learning and Cybernetics,

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*(1-2), 72-87.

Golz, M., Sommer, D., Trutschel, U., Sirois, B., & Edwards, D. (2010). Evaluation of fatigue monitoring technologies. *Somnologie-Schlafforschung und Schlafmedizin, 14*(3), 187-199.

Gratton, G. (2007). Biosignal processing. In J. T. Cacioppo, L. G. Tassinary, G. G. Berntson, J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson (Eds.), *Handbook of psychophysiology (3rd ed.).* (pp. 834-858). Cambridge University Press. https://doi.org/10.1017/CBO9780511546396.035

Guo, M., Li, S., Wang, L., Chai, M., Chen, F., & Wei, Y. (2016). Research on the Relationship between Reaction Ability and Mental State for Online Assessment of Driving Fatigue. *Int J Environ Res Public Health, 13*(12).

Hallvig, D., Anund, A., Fors, C., Kecklund, G., Karlsson, J. G., Wahde, M., & Akerstedt, T. (2013). Sleepy driving on the real road and in the simulator--A comparison. *Accid Anal Prev, 50*, 44-50.

He, J., Liu, D., Wan, Z., & Hu, C. (2014). A noninvasive real-time driving fatigue detection technology based on left prefrontal Attention and Meditation EEG. 2014 International

Conference on Multisensor Fusion and Information Integration for Intelligent Systems (MFI),

Ingre, M., Åkerstedt, T., Peters, B., Anund, A., & Kecklund, G. (2006). Subjective sleepiness, simulated driving performance and blink duration: examining individual differences. *Journal Of Sleep Research, 15*(1), 47-53.

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.

King, L. M., Nguyen, H. T., & Lal, S. K. (2006). Early driver fatigue detection from electroencephalography signals using artificial neural networks. *Conf Proc IEEE Eng Med Biol Soc, 1*, 2187-2190.

Ko, L.-W., Komarov, O., Lai, W.-K., Liang, W.-G., & Jung, T.-P. (2020). Eyeblink recognition improves fatigue prediction from single-channel forehead EEG in a realistic sustained attention task. *Journal of Neural Engineering, 17*(3), 036015.

Lal, S. K. L., & Craig, A. (2001). A critical review of the psychophysiology of driver fatigue. *Biological Psychology, 55*(3), 173-194.

Larue, G. S., Rakotonirainy, A., & Pettitt, A. N. (2011). Driving performance impairments due to hypovigilance on monotonous roads. *Accident Analysis & Prevention, 43*(6), 2037-2046.

Leproult, R., Colecchia, E. F., Berardi, A. M., Stickgold, R., Kosslyn, S. M., & Van Cauter, E. (2003). Individual differences in subjective and objective alertness during sleep deprivation are stable and unrelated. *American Journal Of Physiology. Regulatory, Integrative And Comparative Physiology, 284*(2), 280-290.

Liang, Y., Horrey, W. J., Howard, M. E., Lee, M. L., Anderson, C., Shreeve, M. S., O'Brien, C. S., & Czeisler, C. A. (2019). Prediction of drowsiness events in night shift workers during morning driving. *Accident Analysis and Prevention, 126*, 105-114.

Mahmoodi, M., & Nahvi, A. (2019). Driver drowsiness detection based on classification of surface electromyography features in a driving simulator. *Proc Inst Mech Eng H, 233*(4), 395-406.

Mardi, Z., Ashtiani, S. N. M., & Mikaili, M. (2011). EEG-based drowsiness detection for safe driving using chaotic features and statistical tests. *Journal of Medical Signals and Sensors, 1*(2), 130.

Martensson, H., Keelan, O., & Ahlstrom, C. (2019). Driver Sleepiness Classification Based on Physiological Data and Driving Performance From Real Road Driving. *IEEE Transactions on Intelligent Transportation Systems, 20*(2), 421-430.

Meng, F., Li, S., Cao, L., Li, M., Peng, Q., Wang, C., & Zhang, W. (2015). Driving Fatigue in Professional Drivers: A Survey of Truck and Taxi Drivers. *Traffic Injury Prevention, 16*(5), 474-483.

Miller, K. A., Filtness, A. J., Anund, A., Maynard, S. E., & Pilkington-Cheney, F. (2020). Contributory factors to sleepiness amongst London bus drivers. *Transportation Research Part F: Traffic Psychology and Behaviour, 73*, 415-424.

Min, J., Wang, P., & Hu, J. (2017). Driver fatigue detection through multiple entropy fusion analysis in an EEG-based system. *PLoS ONE, 12*(12), e0188756.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *J Clin Epidemiol, 62*(10), 1006-1012.

National Highway Traffic Safety Administration. (1998). Drowsy driving and automobile crashes [Research Paper].

Parikh, R., Mathai, A., Parikh, S., Chandra Sekhar, G., & Thomas, R. (2008). Understanding and using sensitivity, specificity and predictive values. *Indian journal of ophthalmology, 56*(1), 45-50.

Patel, S., Park, H., Bonato, P., Chan, L., & Rodgers, M. (2012). A review of wearable sensors and systems with application in rehabilitation. *Journal of neuroengineering and rehabilitation, 9*(1), 1-17.

Persson, A., Jonasson, H., Fredriksson, I., Wiklund, U., & Ahlström, C. (2020). Heart Rate Variability for Classification of Alert Versus Sleep Deprived Drivers in Real Road Driving Conditions. *IEEE Transactions on Intelligent Transportation Systems*, 1-10.

Phyo Phyo, S., Sai Ho, L., Rifai, C., Tran, Y., Craig, A., & Hung, N. (2016). EEG-based driver fatigue detection using hybrid deep generic model. *Conf Proc IEEE Eng Med Biol Soc, 2016*, 800-803.

Pritchett, S., Zilberg, E., Xu, Z. M., Karrar, M., Burton, D., & Lal, S. (2011). Comparing accuracy of two algorithms for detecting driver drowsiness — Single source (EEG) and hybrid (EEG and body movement). 7th International Conference on Broadband Communications and Biomedical Applications,

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.

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

Saini, V., & Saini, R. (2014). Driver drowsiness detection system and techniques: a review. *International Journal of Computer Science and Information Technologies, 5*(3), 4245-4249.

Schleicher, R., Galley, N., Briest, S., & Galley, L. (2008). Blinks and saccades as indicators of fatigue in sleepiness warnings: looking tired? *Ergonomics, 51*(7), 982-1010.

Tefft, B. C. (2014). *Prevalence of motor vehicle crashes involving drowsy drivers, United States, 2009-2013*. Citeseer.

Van Dongen, H. P., Baynard, M. D., Maislin, G., & Dinges, D. F. (2004). Systematic interindividual differences in neurobehavioral impairment from sleep loss: Evidence of trait-like differential vulnerability. *Sleep, 27*(3), 423-433.

Van Dongen, H. P., Bender, A. M., & Dinges, D. F. (2012). Systematic individual differences in sleep homeostatic and circadian rhythm contributions to neurobehavioral impairment during sleep deprivation. *Accident Analysis & Prevention, 45, Supplement*(0), 11-16.

Vanlaar, W., Simpson, H., Mayhew, D., & Robertson, R. (2008). Fatigued and drowsy driving: A survey of attitudes, opinions and behaviors. *Journal of Safety Research, 39*, 303-309.

Vicente, J., Laguna, P., Bartra, A., & Bailon, R. (2016). Drowsiness detection using heart rate variability. *Med Biol Eng Comput, 54*(6), 927-937.

Watling, C. N. (2016). *The sleep and wake drives: exploring the genetic and psychophysiological aspects of sleepiness, motivation, and performance* [PhD thesis, Queensland University of Technology]. Brisbane, Australia.

Watling, C. N., Åkerstedt, T., Kecklund, G., & Anund, A. (2016). Do repeated rumble strip hits improve driver alertness? *J Sleep Res, 25*(2), 241-247.

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.

Williamson, A., Lombardi, D. A., Folkard, S., Stutts, J., Courtney, T. K., & Connor, J. L. (2011). The link between fatigue and safety. *Accident Analysis & Prevention, 43*(2), 498-515.

Zeller, R., Williamson, A., & Friswell, R. (2020). The effect of sleep-need and time-on-task on driver fatigue. *Transportation Research Part F: Traffic Psychology and Behaviour, 74*, 15-29.

Zhang, Z., & Zhang, J. (2010). A new real-time eye tracking based on nonlinear unscented Kalman filter for monitoring driver fatigue. *Journal of Control Theory and Applications, 8*(2), 181-188.