Heart Rate Variability for Classification of Alert
Versus Sleep Deprived Drivers in Real
Road Driving Conditions

Anna Persson, Hanna Jonasson, Ingemar Fredriksson, Urban Wiklund, Member, IEEE,
and Christer Ahlström, Senior Member, IEEE

Abstract—Driver sleepiness is a contributing factor in many road fatalities. A long-standing goal in driver state research has therefore been to develop a robust sleepiness detection system. It has been suggested that various heart rate variability (HRV) metrics can be used for driver sleepiness classification. However, since heart rate is modulated not only by sleepiness but also by several other time-varying intra-individual factors such as posture, distress, boredom and relaxation, it is relevant to highlight not only the possibilities but also the difficulties involved in HRV-based driver sleepiness classification. This paper investigates the reliability of HRV as a standalone feature for driver sleepiness detection in a realistic setting. Data from three real-road driving studies were used, including 86 drivers in both alert and sleep-deprived conditions. Subjective ratings based on the Karolinska sleepiness scale (KSS) were used as ground truth when training four binary classifiers (K-nearest neighbours, support vector machine, AdaBoost, and random forest). The best performance was achieved with the random forest classifier with an accuracy of 85%. However, the accuracy dropped to 64% for three-class classification and to 44% for subject-independent, leave-one-participant-out classification. The worst results were obtained in the severely sleepy class. The results show that in realistic driving conditions, subject-independent sleepiness classification based on HRV is poor. The conclusion is that more work is needed to control for the many confounding factors that also influence HRV before it can be used as input to a driver sleepiness detection system.

Index Terms—Driver sleepiness, detection, classification, heart rate variability.

I. INTRODUCTION

Driver sleepiness is a contributing factor to approximately 20% of all road fatalities [1], [2]. Driver sleepiness detection systems may reduce these numbers by convincing sleepy drivers to pull over for a rest or nap. Such systems are typically based on (i) vehicle-based information such as lane keeping performance, (ii) behavioural information such as yawning and eye movements, and/or (iii) driver physiological data such as electrocardiography (ECG), electroencephalography (EEG), or electrooculography (EOG) [3]–[5].

Today’s and tomorrow’s advanced driver assistance and automated driving systems are changing the playing field for sleepiness detection. When longitudinal and lateral positioning is secured by the vehicle, vehicle-based data can no longer be used for sleepiness detection. Similarly, camera-based systems may not see the driver’s face due to obstructed camera views or because the driver is out of position, both of which are more likely during assisted or automated driving. At the same time, physiological data are becoming more available thanks to wearable [6], [7] or remote sensors that can be embedded in the steering wheel, seat belt, and seat [8], [9]. This has revived heart rate and heart rate variability (HRV)-based concepts, and it is logical to use the relationship between cardiac function and sleepiness when designing sleepiness detection systems.

The heart rate is regulated by the sympathetic (activation) and parasympathetic (rest) branches of the autonomic (involuntary) nervous system. Sympathetic activation leads to increased heart rate, more forceful heart contractions, dilated airways, higher respiration rate and increased muscle strength. Parasympathetic activation causes reduced heart rate, decreased blood pressure, and stimulates digestion and waste disposal. Heart rate, and thus HRV, reflects the balance between sympathetic and parasympathetic activity. In general, a lowered heart rate gives more room for variability between successive heartbeats allowing higher HRV. This is typically seen during sleepiness when the body is winding down to prepare for sleep. It is important to note that while a change in alertness level will likely cause a change in HRV (Fig. 1.), the opposite is not necessarily true, as a change in HRV could equally be due to a change in any of the other factors that also affect HRV. These other factors include both inter-individual factors, such as age, gender, and medical conditions, and intra-individual factors, some of which vary over time [10], [11].

It has been hypothesized that parasympathetic activity should increase with increasing levels of driver sleepiness. High-frequency (HF) power in the HRV signal, which is a sign of parasympathetic activity, has indeed been found...
to increase with increasing levels of sleepiness [12], [13]. However, it has also been found to decrease [14] or to display unreliable changes [15]–[17]. It has also been hypothesized that a sleepy driver should have either decreased or increased sympathetic activity, depending on whether or not the driver is struggling to remain awake. Low-frequency (LF) power in the HRV signal, which is often used as an indicator of sympathetic activity, has consequently been found to decrease [12], [14], increase [13], [16], or display unreliable changes [15], [17], [18] with increasing levels of driver sleepiness. These differing results would make sense if increased LF activity was found in experiments with manual driving (due to the struggle to remain awake) while decreased LF activity was found in studies with conditional or full automation (since the driver is then allowed to relax and possibly even sleep). However, all reviewed results come from manual driving experiments. The ambiguous hypotheses are nevertheless very convenient as they provide an explanation regardless of the direction of the results.

For the reasons outlined above, it is relevant to question how appropriate it is to use HRV for driver sleepiness detection outside a controlled experimental environment. The primary aim of this paper is therefore to investigate whether HRV metrics alone, measured during real-road driving, can be used for driver sleepiness detection. Preliminary results of this study were presented by Persson, et al. [19].

II. MATERIALS AND METHODS

Two methodological approaches have been used for the HRV analyses in this paper: a group-level statistical approach and a machine learning approach. Common to both approaches was a pre-processing stage in which the ECG signals were filtered and divided into 5-min epochs. The HRV metrics (or features) under investigation were then calculated for each epoch. The first methodological approach involved an analysis of variance (ANOVA). This was done to see whether the dataset provided results similar to those previously reported in the HRV literature. The second methodological approach used a standard machine learning pipeline to investigate how appropriate it is to design a sleepiness classifier based on HRV metrics alone.

A. Sleepiness Database

The database used in this paper consists of data from three separate driver sleepiness experiments (see [20]–[22] for a detailed account). To the best of our knowledge, this is one of the largest labelled real-road driver sleepiness datasets with data from both alert and sleep-deprived conditions. Experiment 1 included 18 drivers (8 women, mean age 41 ± 9 years) who drove for about 90 min on a motorway in real traffic. Each driver drove two times, once in a supposedly alert state during daytime and once in a sleep-deprived state during night-time. Experiment 2 included 24 drivers (12 women, mean age 35 ± 10 years) who drove for about 135 min three times on a motorway (supposedly alert during daytime, mostly alert in the evening, and sleep deprived during night-time). Experiment 3 included 44 drivers (21 women, mean age 45 ± 8 years) who drove for about 90 min three times on a rural road (supposedly alert during daytime, mostly alert in the evening, and sleep deprived during night-time). The participants were recruited by random selection from the Swedish register of vehicle owners. All participants were prepared in the same way in all experiments. Before arrival, the participants were requested to avoid alcohol for 72 h and to abstain from nicotine and caffeine for 3 h before driving. All participants reported that they were healthy with good to excellent sleep quality. The Swedish government approved the testing of sleepy drivers on real roads (N2007/5326/TR), and each of the three experiments was approved by the Regional Ethics Committee in Linköping, Sweden. The experimental car was equipped with dual command, allowing the test leader to take control of the vehicle if necessary.

The ECG (lead II) was measured using disposable Ag/AgCl electrodes connected to a portable digital recording system (Vitaport 2 and 3, Temec Instruments BV, the Netherlands) using a sampling rate of 256 Hz. The Karolinska Sleepiness

Fig. 1. Model of how different time-varying intra-individual factors affect the cardiovascular control system. The autonomic nervous system controls heart rate via a complex feedback system, in which homeostasis is maintained by information received by the baroreceptors acting on changes in arterial blood pressure. Sleepiness is but one of the many time-varying intra-individual factors that modulate heart rate and HRV.
Scale (KSS) was used to acquire self-reported sleepiness every fifth minute. KSS has nine levels [23]: 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. The reported value corresponds to the average feeling over the past 5 min.

The KSS values are used as target values when training the classifiers. Alternative approaches used as ground truth in driver sleepiness studies include EEG [24], [25], reaction time tests [26], and expert ratings based on observations [12], [15]. However, reaction time tests are difficult to administer in real-road driving and video-based expert ratings have been found to be unreliable [27]. EEG-based measures suffer from noise in naturalistic settings, large inter-individual variability, and the fact that some individuals do not respond despite being clearly sleepy [3], [28], [29]. The main drawbacks of KSS are that the subjective feeling does not always reflect the actual sleepiness level [30], repeated reporting can have an alerting effect [31], and participants may interpret the levels of KSS differently. Advantages are that KSS correlates with lane departures, is easily applied, and has been found to be the measure of driver sleepiness least affected by inter-individual variations [23]. All in all, subjective ratings seem to be the best option.

B. Pre-Processing

R-peaks were extracted from the ECG using the Pan Tompkins algorithm [32] and the RR time series were extracted as the time difference between heart beats. The corresponding normal to normal (NN) time series were obtained by removing outliers using the standard deviation method [33], [34]. Here the threshold was set to 4 standard deviations. Epochs with more than 5% outliers were discarded. In total, five complete recordings were discarded due to poor signal quality, and the outlier removal step removed 13 additional 5-min epochs.

C. HRV Feature Extraction

All NN signals were divided into 5-min epochs, partly because the KSS values were reported every fifth minute, but also because this is the recommended minimum duration for short-term HRV analysis [35]. This resulted in 3954 epochs.

Twenty-four commonly used HRV features were extracted according to Table I. These 24 features make up the main feature set. This feature set, which will be referred to as the HRV feature set. Due to the large individual differences encountered in HRV analyses [10], [11], [17], [36], a second baseline-corrected feature set was constructed by subtracting the mean feature values per participant corresponding to the KSS ≤ 5 cases. This will be called the baseline-corrected feature set. Finally, the time difference between consecutive 5-min epochs of the 24 original features was computed. This feature set will be called the time-difference feature set. All 72 features were standardized by removing the mean and dividing by the standard deviation.

The frequency domain features were derived from an autoregressive power spectral density estimate using the modified covariance method with a model order of 32. The VLF band was defined as 0.003–0.04 Hz, the LF band as 0.04–0.15 Hz, and the HF band as 0.15–0.4 Hz. Sample entropy was derived using embedding dimension = 2 and threshold = 0.2 × std(NN) according to. Richman and Moorman [37]. The potentials of unbalanced complex kinetics (PUCK) metrics were implemented according to. Igasaki, et al. [18].

D. Statistical Analysis

The HRV metrics were analysed with mixed-model ANOVAs using two fixed factors: Daytime (day or evening versus night-time driving) and Sleepiness (KSS ≤ 5, KSS = 6, KSS = 7, KSS = 8, and KSS = 9). Participant (1–86) was included as a random factor. The key effect of interest was how different HRV metrics relate to subjective sleepiness. The factors Daytime and Sleepiness are dependent, but since high KSS levels are also common during the day, both factors were included in the analysis. The other two factors should be considered confounding factors. The significance level was set to 0.01, which corresponds to \( p < 0.0004 \), with Bonferroni correction to compensate for the 24 comparisons.

E. Sleepiness Classification

An overview of the machine learning pipeline is illustrated in Fig. 2. The feature set was split into three parts: 30% for feature selection and parameter tuning, 50% for training, and 20% for testing. To make sure that the results are not just coincidental, stemming from a certain data partitioning, this process was repeated ten times. The final test results are mean values across these ten repetitions. This 10-fold cross validation will be referred to as classification-repetitions.

Three classes of subjective sleepiness were defined: alert (KSS ≤ 5), somewhat sleepy (6 ≤ KSS ≤ 7), and severely sleepy (KSS ≥ 8). In the case of binary classification, the somewhat sleepy class was left out to obtain a clearer distinction between the alert and sleepy classes [38]. The class definitions are justified by the observation that hardly any line crossings occur at KSS ≤ 5, whereas a markedly increased frequency of unintentional line crossings occurs at KSS ≥ 8 [22].

Feature selection was carried out using sequential forward floating selection (SFFS) [39]. SFFS was wrapped with a binary decision tree classifier, 5-fold cross-validation, 20 cross-validation runs, and a trade-off between sensitivity and specificity as optimization score. Since SFFS often results in low-dimensional non-redundant but noise sensitive feature sets, the SFFS procedure was run repeatedly (20 times) on different partitions of the feature selection set. The features selected in ≥20% of the repetitions were used as the final feature set.

Four different binary classifiers were evaluated: \( k \)-nearest neighbours (kNN), support vector machine (SVM), AdaBoost, and random forest. These classifiers were chosen because they are well established and because there is a clear difference in complexity and computational cost between them. The \( k \)NN used 25 neighbours and a Euclidean distance function with no distance weighting. The SVM used a Gaussian kernel, a heuristic procedure to set an appropriate kernel scale factor,
and a box constraint level $=$ 10. The AdaBoost classifier used decision trees as weak learners (50 voting trees) with the maximum number of decision splits set to 20 and a learning rate of 0.1. The AdaBoost.M1 algorithm was used as the ensemble aggregation method. The random forest also used decision trees as weak learners (50 voting trees) and applied bootstrap aggregation as the ensemble-aggregation method. A cost function was used to avoid misclassifications of the severely sleepy class, since the dataset was unbalanced (54% alert, 32% somewhat sleepy, and 14% severely sleepy). The number of neighbours, box constraint level, and number of trees, respectively, were selected by evaluating classification performance over a range of parameter values. Except for these parameters, the default settings in the MATLAB Statistics and Machine Learning Toolbox version 11.1 (The Mathworks Inc., Natick, MA, USA) were used.
The best performing binary classifier was also evaluated in a three-class classification setting. This was done using both 10-fold cross-validation as outlined above and leave-one-participant-out (LOO) cross-validation. The LOO evaluation was added to investigate subject-independent performance.

All classification results are derived based on three different feature sets: (i) the HRV feature set, (ii) the HRV feature set combined with the baseline-corrected feature set, and (iii) all three feature sets combined. This was done to investigate the importance of a personalized feature set and to investigate the added value of taking time history into account [17].

### III. RESULTS

The ANOVA results are summarized in Table II and Fig. 3. Almost all HRV metrics varied with the level of driver sleepiness, except for the absolute power in VLF, the LF peak, the absolute power in HF, sample entropy, and the slope from the PUCK analysis. By and large, the results support the hypothesis that sleepiness is associated with lowered heart rate, more HRV, increased LF power (i.e. increased sympathetic activity – fighting to stay awake), and increased LF power (i.e. increased vagal activity). Note that the relative power in HF is changing in the “wrong” direction, which is a consequence of the comparatively stronger influence of LF when normalizing with LF+HF. Also note that the baseline-normalized relative power in HF does not show this reversed behaviour. The random factor participant was significant for all HRV metrics, suggesting large inter-individual differences.

Many of these features represent similar information (Fig. 3) and it is likely that many features are redundant. It is therefore warranted to use feature selection to reduce the complexity of the developed classifier. The SFFS iterations show that both the number and selection of features varied substantially from iteration to iteration (Fig. 4). Five features were selected in all iterations: RMSSD, NN50, pNN50, meanNN, and meanHR from the PUCK analysis. When adding the baseline-corrected feature set, many of these features were replaced by their baseline-corrected counterparts (e.g. RMSSD, meanNN, absolute power in LF, and SSD1 from the PUCK analysis). When adding the time-difference feature set changed the result only marginally.

The best overall binary classification performance was achieved with the random forest classifier, but both kNN and AdaBoost had higher sensitivity (Table III). The mean...
TABLE III

| Test | HRV feature set, mean ± SD (range) | HRV + baseline corrected | HRV+ baseline corrected + time difference |
|------|-----------------------------------|--------------------------|------------------------------------------|
| **Accuracy** | kNN | 67.2 ± 3.5 (59.6–71.1) | 77.5 ± 1.6 (75.3–80.1) | 77.9 ± 3.8 (71.9–85.0) |
| | SVM | 79.0 ± 2.4 (72.6–81.5) | 83.4 ± 2.0 (80.2–86.9) | 79.9 ± 4.5 (71.1–84.9) |
| | AdaBoost | 75.1 ± 1.9 (71.0–77.6) | 82.4 ± 1.4 (80.4–84.4) | 82.1 ± 2.3 (78.3–85.4) |
| | Random Forest | 80.9 ± 1.8 (76.5–82.8) | 85.4 ± 2.1 (81.9–88.7) | 84.3 ± 2.3 (81.2–87.8) |
| **Sensitivity** | kNN | 69.6 ± 7.8 (59.3–79.6) | 72.4 ± 6.2 (59.8–82.9) | 74.3 ± 6.7 (62.3–82.6) |
| | SVM | 43.0 ± 10.5 (26.5–55.7) | 50.2 ± 12.6 (19.8–61.5) | 48.3 ± 11.9 (29.6–69.4) |
| | AdaBoost | 61.6 ± 4.7 (56.4–72.5) | 70.4 ± 5.0 (60.6–76.1) | 69.8 ± 4.5 (59.4–74.6) |
| | Random Forest | 53.5 ± 4.8 (45.9–60.8) | 66.9 ± 5.4 (55.9–73.4) | 65.9 ± 7.6 (50.0–78.3) |
| **Specificity** | kNN | 66.6 ± 5.8 (54.4–73.2) | 79.0 ± 2.4 (73.3–82.2) | 78.9 ± 5.6 (70.5–88.2) |
| | SVM | 88.6 ± 5.3 (77.5–96.1) | 92.3 ± 3.9 (86.3–98.4) | 88.4 ± 5.7 (75.6–95.6) |
| | AdaBoost | 78.7 ± 2.8 (75.0–82.8) | 85.7 ± 1.7 (82.9–98.2) | 85.4 ± 2.8 (81.2–98.4) |
| | Random Forest | 88.2 ± 1.5 (85.4–90.9) | 90.4 ± 1.4 (88.5–93.4) | 89.3 ± 2.2 (86.5–92.9) |
| **F1** | kNN | 47.0 ± 2.0 (44.5–50.0) | 57.6 ± 3.4 (53.9–63.7) | 59.1 ± 5.1 (51.6–69.3) |
| | SVM | 45.6 ± 4.7 (37.6–53.0) | 55.3 ± 8.9 (31.4–63.3) | 50.4 ± 9.8 (33.8–63.4) |
| | AdaBoost | 50.8 ± 2.5 (45.7–54.7) | 62.8 ± 3.2 (59.5–68.2) | 62.4 ± 4.3 (57.5–69.1) |
| | Random Forest | 53.9 ± 3.1 (46.9–57.3) | 65.9 ± 4.8 (59.4–75.5) | 64.1 ± 0.5 (51.5–72.6) |

TABLE IV

| Test | HRV feature set Mean ± SD (range) | HRV + baseline corrected | HRV+ baseline corrected + time difference |
|------|-----------------------------------|--------------------------|------------------------------------------|
| **Mean of classes** | Accuracy | 57.9 ± 1.4 (55.8–59.9) | 64.2 ± 1.5 (61.1–66.0) | 62.2 ± 2.1 (58.7–65.1) |
| | Sensitivity | 49.7 ± 1.1 (47.5–51.1) | 57.9 ± 2.0 (53.8–60.5) | 55.5 ± 3.1 (50.1–60.8) |
| | Specificity | 75.5 ± 0.7 (74.1–76.5) | 79.3 ± 1.0 (77.5–80.4) | 78.4 ± 1.3 (75.8–80.8) |
| | F1 | 49.0 ± 1.2 (46.5–51.3) | 56.9 ± 1.9 (53.7–59.7) | 54.2 ± 3.1 (49.1–58.7) |
| **Per class** | Sensitivity “severe sleepiness” | 39.7 ± 2.6 (36.9–44.1) | 53.7 ± 4.3 (43.3–58.1) | 50.6 ± 7.4 (43.6–62.5) |
| | Sensitivity “alert” | 78.5 ± 2.4 (73.5–81.7) | 83.3 ± 1.3 (81.9–86.3) | 82.7 ± 2.3 (78.9–85.2) |
| | Sensitivity “somewhat sleepy” | 31.0 ± 2.4 (27.2–36.6) | 36.6 ± 2.2 (31.8–39.7) | 33.2 ± 4.6 (26.4–40.6) |
| | Specificity “severe sleepiness” | 87.8 ± 1.5 (85.2–90.8) | 89.2 ± 1.3 (85.3–91.6) | 87.5 ± 2.1 (85.1–92.2) |
| | Specificity “alert” | 54.7 ± 2.8 (48.9–57.8) | 61.9 ± 3.0 (56.3–65.2) | 61.1 ± 3.4 (54.5–68.3) |
| | Specificity “somewhat sleepy” | 84.1 ± 1.9 (79.6–85.9) | 86.8 ± 1.1 (85.7–89.4) | 86.8 ± 1.4 (84.9–89.6) |
| | F1 “severe sleepiness” | 37.4 ± 2.5 (34.3–42.3) | 49.3 ± 3.5 (41.5–54.6) | 45.0 ± 5.5 (35.4–52.8) |
| | F1 “alert” | 73.3 ± 1.7 (69.3–74.9) | 77.1 ± 1.5 (73.8–78.8) | 76.4 ± 1.6 (73.7–79.2) |
| | F1 “somewhat sleepy” | 37.5 ± 1.8 (34.4–41.2) | 44.4 ± 2.3 (39.5–47.4) | 41.1 ± 5.1 (33.2–47.9) |

Accuracy of the random forest classifier was 81% when the HRV feature set was used as input. This increased to 85% when including the baseline-corrected features. Adding time history did not improve the results.

Performance dropped by about 20 percentage points for multi-class classification (Table IV). Despite the cost function, only a small share of the severely sleepy drivers was identified as sleepy. To some extent (an additional 5 percentage points), this was remedied by adding the baseline-corrected features. Compared with the multi-class cross-validation results, the LOO results dropped by another 15–20 percentage points (Table IV). This suggests that personalized algorithms, not just personalized features, are needed in a driver sleepiness monitoring system based on HRV.

IV. DISCUSSION

The suitability of using HRV metrics for driver sleepiness classification has been evaluated. Significant effects of sleepiness were found on most HRV metrics and the binary classification results reached an accuracy of 85.4%. However, the accuracy dropped by about 20 percentage points for three-class classification, and by almost another 20 percentage points for subject-independent LOO classification. Sensitivity was low in general, with the worst results in the severely sleepy drivers.
Driver fatigue is defined as a suboptimal psychophysiological condition caused by exertion [40]. Sleep-related fatigue, i.e., sleepiness or drowsiness, is defined as a “a physiological drive to fall asleep” [41] caused by sleep loss, circadian rhythm and time since awakening. Certain characteristics of driving, like task demand and driving environment, can produce task-related fatigue in the absence of any sleep-related cause. Task-related fatigue can then also be subdivided as fatigue due to overload and underload, respectively [42]. Long-term fatigue that isn’t relieved by sleep, such as chronic fatigue syndrome, and muscle fatigue, such as after sports, are out of scope of this paper. A summary of previous research on driver sleepiness/fatigue and HRV is presented in Table V.

Note that differences in experimental design, state inducement, and validation approaches make it difficult to compare results across studies. The studies in Table V should thus only be compared on a higher level. Here it can be seen that Patel, et al. [43] reached an accuracy of 90% (sensitivity not reported) with a neural network classifier. Li and Chung [44] obtained an accuracy of 95% (sensitivity = 95%) with wavelet analysis and an SVM classifier. Fujiwara, et al. [45] used multivariate statistical process control and managed to predict 12 out of 13 sleep onsets. Abe, et al. [36] managed to predict 7 out of 8 sleepiness-related accidents. Finally, Vicente, et al. [17] achieved a positive predictive value of 96% (sensitivity = 59%) with a linear discriminant analysis classifier. Our achieved accuracy is lower than those reported in these other studies. However, our lower sensitivity scores are in line with Vicente, et al. [17]. One reason for the discrepancies could be the reduced level of experimental control in our real-road data as compared with the driving simulator data used in most other studies [12], [18], [36], [43]–[45]. If sleepiness is the only factor manipulated in the experiment while other influencing factors are kept as constant as possible, then a discovered change in HRV is likely due to the altered alertness level. In a real-road setting, the experimental control is weakened, and the association between HRV and sleepiness is weakened. Another reason for the discrepancies could be that the number of participants in the other studies was sometimes as low as 4–12 drivers [15], [18], [43], [44]. A small dataset may be justified in a within-subject design experiment, but it does not allow for proper validation of the sophisticated machine learning methods that are often used. A third reason for the discrepancies could be that different ground truths for sleepiness/fatigue have been used in different studies.

The conclusions drawn in the studies outlined in Table V are that HRV-based sleepiness detection “may add significant improvements to existing car safety systems” [17], “can be used as a fatigue countermeasure” [43], and could “contribute to prevent[ing] accidents caused by drowsy driving” [36], [45]. We draw a different conclusion, namely that it will be difficult
to use HRV for sleepiness detection in a production vehicle. As mentioned in the introduction (see Fig. 1), there are many time-varying intra-individual factors that can cause a change in HRV. Stating that a particular change in HRV is due to sleepiness alone and not to any of these other factors is difficult, especially since other mental states such as boredom and relaxation give rise to very similar changes in the HRV metrics.

There is considerable individual variability in HRV and changes in HRV are different in every person [46]. To circumvent this issue, the use of personalised algorithms [36], [45] or individually normalized features [17] has been suggested. In this study, the latter approach was used when creating the baseline-corrected feature set, something that led to a 5-percentage-point increase in classification accuracy. While personalised algorithms or features may account for inter-individual variability due to age, gender, and other factors, this solution will not be able to account for time-varying intra-individual variability that arises due to distress, anger, boredom, relaxation, etc. [10], [11]. Thus, although some studies have shown promising results for drowsiness and sleepiness detection based on personalised HRV-based algorithms, it is unlikely to be possible to develop a sleepiness detection system based solely on HRV. Multimodal systems that use HRV as one of several inputs have therefore been suggested [47], [48]. However, when doing so, it has been found that HRV does not contribute much when used in combination with EOG, EEG, or vehicle measures [47], [49].

There are some inherent issues with HRV that need to be accounted for even when personalised algorithms are used. One challenge is that diseases such as diabetes mellitus markedly reduce HRV to a stable but very low magnitude [50], which in turn makes it very difficult or even impossible to detect changes over time. The same is true for high age [51], when abnormal HRV patterns have been observed that could mask changes due to sleepiness observed with advancing

| Study | N  | Sim/real road | State induction | Ground truth | Features | Analysis/classifier | Result | Conclusion |
|-------|----|---------------|-----------------|--------------|----------|---------------------|--------|------------|
| [12]  | 11 | Sim           | Monotonous driving without sleep deprivation | Video-based observer ratings | LF, HF, LF/HF | Paired t-test | Significant differences between conditions | "The study shows very encouraging results that can be used to prevent drowsiness related accidents". |
| [15]  | 10 | Road          | Monotonous driving without sleep deprivation | Video-based observer ratings | 11 features | Paired t-test | Significant differences in 8 of 11 HRV metrics | "Results suggest that a drowsiness alarm based on HRV indexes is feasible but cannot rely on only one index". |
| [18]  | 8  | Sim           | Monotonous driving without sleep deprivation | Subjective self-ratings, PUCK | LF, HF, LF/HF | ANOVA | Significant statistical differences in almost all subjects | "PUCK analysis of heart rate variability may be useful for assessing drowsiness while driving". |
| [34]  | 76 | Road          | Sleep deprivation | Subjective self-ratings | 12 personalized features | ANOVA | Significant differences between conditions in all HRV metrics | "HRV analysis shows promise for driver sleepiness detection". |
| [17]  | 30 | Sim and road  | Sleep deprivation or monotonous driving without sleep deprivation | Video-based observer ratings | 96 features | Stepwise linear discriminant analysis | Leave one [participant] out; positive predictive value = 0.96, sensitivity = 0.59, specificity = 0.98 | "Incorporating drowsiness assessment based on HRV signal may add significant improvements to existing car safety systems". |
| [36]  | 27 | Sim           | Monotonous driving without sleep deprivation | Video-based observer ratings | LF, HF, LF/HF, SDNN, RMSSD, meanNN, NN50, total power | Multivariate statistical process control | Predicted 7 of 8 cases of drowsiness before driving accidents occurred | "The possibility of realizing an HRV-based drowsy driving accident prediction system was demonstrated through the driving simulator experiments". |
| [45]  | 34 | Sim           | Monotonous driving without sleep deprivation | Sleep onset based on EEG | LF, HF, LF/HF, SDNN, RMSSD, meanNN, NN50, total power | Multivariate statistical process control | Predicted 12 of 13 sleep onsets, with a false positive rate of 1.7 times per hour. | "The proposed HRV-based drowsy driving detection algorithm is more promising than other conventional methods with respect to accuracy as well as practical use". |
| [44]  | 4  | Sim           | Alert versus drowsy state; unclear how they were induced | Percentage eye closure | Entropy, variance, kurtosis, multiscale component | SVM | Leave one [observation] out; accuracy = 0.95, specificity = 0.95. | "Results indicate that a better real-time driver drowsiness detection system can be developed by using wavelet-based features". |
| [43]  | 12 | Sim           | Monotonous driving with slight sleep deprivation | Video-based observer ratings | Spectral image (30 x 30) | Feed-forward neural network | Hold out accuracy = 0.90. | "This HRV-based fatigue detection technique can be used as a fatigue countermeasure". |
have treated heart conditions or diabetes mellitus that do not
pain/tachycardia/respiratory problems, and had good to excel-
participants. Via the background questionnaire, the participants
detector. A limitation of the dataset used in this study is
since the extrasystolic beats mask the underlying autonomic
HRV analysis [17]. However, sleepiness cannot be evaluated in
the time window when they occur can be excluded from the
heart rate. Spurious extrasystolic beats can be detected and
they increase the magnitude of the beat-to-beat variation in
extrasystolic beats must be detected and accounted for since
age [52]. HRV is also severely affected by arrhythmia, and
extrasytolic beats cannot be detected based on HRV,
since the extrasystolic beats mask the underlying autonomic
modulation of heart rate. This means that a proportion of
the population will never be able to use an HRV-based sleepiness
detector. A limitation of the dataset used in this study is
that we have limited knowledge of the health status of the
participants. Via the background questionnaire, the participants
reported that they were of good health, did not have chest
pain/tachycardia/respiratory problems, and had good to excel-
ent sleep quality. However, we do not know whether they
have treated heart conditions or diabetes mellitus that do not
affect their general health.

V. CONCLUSION

Partial or conditional automation will transform the role of
the driver from active driving to passive monitoring. This
will lead to increased levels of fatigue due to boredom and
under-stimulation [53], and consequently an increased preva-
ence of driver sleepiness and fatigue. Finding robust driver
monitoring solutions is therefore increasingly important.

Generalising an HRV-based sleepiness detector is a major
challenge, as HRV data vary both between individuals and
over time within individuals, depending on both internal and
external factors. A successful classifier should thus make use
of driver profiles that evolve over time [37]. Future research
on HRV-based driver sleepiness characterization should focus
on complementing personalised algorithms with functionalities
able to account and control for the many confounding factors
known to affect HRV.

ACKNOWLEDGMENT

The authors gratefully acknowledge T. Åkerstedt at the
Karolinska Institute and A. Anund at VTI for their valuable
comments on an early draft of this manuscript.

REFERENCES

[1] D. Zwahlen, C. Jackowski, and M. Pfäffli, "Sleepiness, driving, and
motor vehicle accidents: A questionnaire-based survey," J. Forensic
Legal Med., vol. 44, pp. 183–187, Nov. 2016.
[2] K. A. Armstrong, P. Obst, T. Banks, and S. S. Smith, "Managing driver
fatigue: Education or motivation?" Road Transp. Res., vol. 19, no. 3,
p. 14–20, 2010.
[3] A. R. Sparrow, C. M. LaJambre, and H. P. A. Van Dongen, "Drowsiness
measures for commercial motor vehicle operations," Accident Anal.
Prevention, vol. 126, pp. 146–159, May 2019.
[4] M. I. Chacon-Murguia and C. Prieto-Resendiz, "Detecting driver drowsi-
ness: A survey of system designs and technology," IEEE Consum.
Electron. Mag., vol. 4, no. 4, pp. 107–119, Oct. 2015.
[5] G. Sikander and S. Anwar, "Driver fatigue detection systems: A review," IEEE Trans. Intell. Transp. Syst., vol. 20, no. 6, pp. 2339–2352,
Jun. 2019.
[6] H. Lee, J. Lee, and M. Shin, "Using wearable ECG/PPG sensors for
driver drowsiness detection based on distinguishable pattern of
recurrence plots," Electronics, vol. 8, no. 2, p. 192, 2019.
[7] S. Seneviratne et al., "A survey of wearable devices and chal-
enges," IEEE Commun. Surveys Tuts., vol. 19, no. 4, pp. 2573–2620,
4th Quart., 2017.
[8] S. Leonhardt, L. Leicht, and D. Teichmann, “Unobtrusive vital sign
monitoring in automotive environments—A review,” Sensors, vol. 18,
no. 9, p. 3080, 2018.
[9] J. Pinto, J. Cardoso, A. Lourengão, and C. Carreiras, “Towards a
continuous biometric system based on ECG signals acquired on the
steering wheel,” Sensors, vol. 17, no. 10, p. 2228, 2017.
[10] M. Lohani, R. B. Payne, and D. L. Strayer, “A review of psychophys-
iological measures to assess cognitive states in real-world driving,”
Frontiers Hum. Neurosci., vol. 13, p. 57, Mar. 2019.
[11] S. Laborde, E. Mosley, and J. F. Thayer, “Heart rate variability
and cardiac vagal tone in psychophysiological research—recommenda-
tions for experiment planning, data analysis, and data reporting,” Frontiers
Psychol., vol. 8, p. 213, Feb. 2017.
[12] M. Awaiss, N. Badruddin, and M. Drieberg, “A non-invasive approach to
detect drowsiness in a monotonous driving environment,” in Proc. IEEE
Region Conf. (TENCON), Oct. 2014, pp. 1–4.
[13] J. Berg, N. Geely, U. Wiklund, and U. Landstrom, “Heart rate variability
during sedentary work and sleep in normal and sleep-deprived states,” Clin.
Physiol. Funct. Imag., vol. 25, no. 1, pp. 51–57, Jan. 2005.
[14] R. J. Apparies, T. C. Ritsma, and S. W. Porges, “A psychophysiological
investigation of the effects of driving longer-combination vehicles,”
Ergonomics, vol. 41, no. 5, pp. 581–592, May 1998.
[15] N. Rodriguez-Ibañez, M. A. García-Gonzalez, M. A. F. de la Cruz,
M. Fernández-Chimeno, and J. Ramos-Castro, “Changes in heart rate
variability indexes due to drowsiness in professional drivers measured
in a real environment,” in Proc. Comput. Cardiol., vol. 39, Sep. 2012,
p. 913–916.
[16] Y. Tran, N. Wijesuriya, M. Tarvainen, P. Karjalainen, and A. Craig,"The relationship between spectral changes in heart rate variability and
fatigue," J. Psychophysiol., vol. 23, no. 3, pp. 143–151, Jan. 2009.
[17] J. Vicente, P. Laguna, A. Bartra, and R. Bailón, “Drowsiness detection
using heart rate variability,” Med. Biol. Eng. Comput., vol. 54, no. 6,
p. 927–937, Jun. 2016.
[18] T. Igasaki, T. Morie, M. Utsunomiya, I. A. Akbar, and A. M. Rumagui,
“Can potentials of unbalanced complex kinetics of heart rate variability
estimate drowsiness?” in Proc. 10th Biomed. Eng. Int. Conf. (BME-
ICON), Aug. 2017, pp. 1–5.
[19] A. Persson, H. Jonasson, I. Fredriksson, U. Wiklund, and C. Ahlström,
“Heart rate variability for driver sleepiness classification in real road
driving conditions,” in Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol.
Soc. (EMBC), Berlin, Germany, Jul. 2019, pp. 6537–6540.
[20] A. Anund, C. Fors, D. Hallvig, T. Åkerstedt, and G. Kecklud, “Observer
rated sleepiness and real road driving: An explorative study,” PLoS ONE,
vol. 8, no. 5, 2013, Art. no. e64782.
[21] C. Fors, C. Ahlström, and A. Anund, “A comparison of driver sleepiness
in the simulator and on the real road,” J. Transp. Saf. Secur., vol. 10,
os. 1–2, pp. 72–87, Mar. 2018.
[22] D. Hallvig, A. Anund, C. Fors, G. Kecklund, and T. Åkerstedt, “Real
driving at night - Predicting lane departures from physiological and
subjective sleepiness,” Biol. Psychol., vol. 101, pp. 18–23, Sep. 2014.
[23] T. Åkerstedt, A. Anund, J. Axelsson, and G. Kecklund, “Subjective
sleepiness is a sensitive indicator of insufficient sleep and impaired
waking function,” J. Sleep Res., vol. 23, no. 3, pp. 242–254, Jun. 2014.
[24] M. Gillberg, G. Kecklund, and T. Åkerstedt, “Sleepiness and per-
formance of professional drivers in a truck simulator—Comparisons between day and night driving,” J. Sleep Res., vol. 5, no. 1, pp. 12–15, Mar. 1996.
[25] A. Picot, S. Charbonnier, and A. Caplier, “On-line detection of drowsi-
ness using brain and visual information,” IEEE Trans. Syst., Man,
Cybern.-A, Syst. Hum., vol. 42, no. 3, pp. 764–775, May 2012.
[26] Q. Massoz, J. Verly, and M. van Droogenbroeck, “Multi-timescale
drowsiness characterization based on a video of a driver’s face,” Sensors,
vol. 18, no. 9, p. 2901, 2018.
[27] C. Ahlström, C. Fors, A. Anund, and D. Hallvig, “Video-based observer
rated sleepiness versus self-reported subjective sleepiness in real road
driving,” Eur. Transp. Res. Rev., vol. 7, no. 4, p. 38, Dec. 2015.
[28] D. Sandberg et al., “The characteristics of sleepiness during real driving
at night—A study of driving performance, physiology and subjective
experience,” Sleep, vol. 34, no. 10, pp. 1517–1525, 2011.
[29] T. Åkerstedt et al., “Relation of sleepiness to partial sleep deprevation,
time of day and time on task in a driving simulator—the DROWSI
project,” J. Sleep Res., vol. 19, no. 2, pp. 298–309, 2010.
