Learning to Detect Drowsiness in Drivers

Antti Takalahti

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty		Laitos — Institution — Department	
Faculty of Science		Department of Computer Science	
Tekijä — Författare — Author			
Antti Takalahti			
Työn nimi — Arbetets titel — Title			
Learning to Detect Drowsiness in Drivers			
Oppiaine — Läroämne — Subject			
Computer Science			
Työn laji — Arbetets art — Level	Aika — Datum — Month and year		Sivumäärä — Sidoantal — Number of pages
Master's Thesis	April 19, 2016		61
Tiivistelmä — Referat — Abstract			

Supervised machine learning consists of inferring a function from labelled examples. The examples used in this study are twenty second stretches of driving data and the labels are binary values of visually scored drowsiness.The ultimate goal is to predict future driving performance by looking only at the steering wheel angle and the position of the accelerator pedal over time.

Chapter two explains what drowsiness is, how tools such as electroencephalogram and electrooculogram reveal drowsiness, and what other tools have been used to study drowsiness.

Chapter three shows how a driver's drowsiness can be detected, and what the psychological aspects relevant to driving performance are. Some currently used methods are also described.

Chapter four shows how the Karolinska drowsiness score is derived from electroencephalogram and electrooculogram, and how this score is used to train the nearest neighbour classifier to detect drowsiness from individual segments of driving data and further to predict driving performance.

Avainsanat — Nyckelord — Keywords	
Drowsiness, Machine learning, Driving	
Säilytyspaikka — Förvaringsställe — Where deposited	
Muita tietoja — Övriga uppgifter — Additional information	

Contents

1 Introduction

There are more than a billion cars on the roads worldwide, and falling asleep while driving is common [65]. In a study conducted in the United States by the Centers for Disease Control and Prevention, one in every twenty-five adult respondents reported having fallen asleep while driving during the last thirty days [27]. In another study, done in Finland by the Finnish Road Safety Council Liikenneturva, it was shown that a quarter of the population had at least once fallen asleep while driving [49].

The aim of my study is to show how a driver´s drowsiness can be detected using common machine learning methods. I will do this by applying three frequently used methods to a data set collected in a study made in Sweden. In that study, drowsiness was measured in eighteen drivers as they drove for approximately ninety minutes first during daytime and then during the night. The drivers' level of drowsiness was measured with the use of an eye camera, by measuring the electrical activity of the brain and by recording the eye movement.

Machine learning has been previously used to detect sleep from the electrical activity of the brain and in how drivers drive in a simulator by recording how they negotiated curves and then using machine learning to learn the difference in alert and drowsy driving. Being tired changes the behaviour of the driver. While the exact effects of drowsiness are difficult to define, one can use machine learning as long as one has some form of biomarker for the level of drowsiness.

Machine learning has huge potential as the race to develop autonomously driving cars rages on, producing vast amounts of data each day. Soon we will have similar data at our use as airline accident investigators, with cars carefully recording the moments leading up to the accident. Driving performance goals like lane position, distance to other cars and flow of traffic are perfectly suited for machine learning algorithms and learning is a perfect tool for creating solutions to these problems.

I will first look at how drowsiness affects people in general and how it can be measured. I will then take a look at how drowsiness can be detected in drivers and show how machine learning can be used as a tool for drowsiness detection. Finally, I will present my own study, where I try to predict which drivers fail to complete their night time drive by learning from a human

scored measure of drowsiness and applying that knowledge to the steering wheel angle, vehicle speed, and accelerator pedal position.

2 Using Biomarkers to Measure Drowsiness

2.1 Drowsiness, Sleepiness, Tiredness, and Fatigue

In this study, the terms drowsiness, sleepiness, tiredness, and fatigue are used as synonyms. This is of course a gross generalisation, but serves the purposes of this study. The main aim of the study is not to distinguish between these different phenomena, but rather on studying their common effects and to show how they can be measured.

Sleepiness research is usually about problems related to sleep – the effects of drugs and illnesses. A condition where person feels compelled to nap even after a normal night's sleep. Tiredness refers to a condition that causes continuous feeling to tiredness or fatigue. This is generally more about the lack of energy and not about feeling the need to sleep. Fatigue research studies how physical exercise, emotional stress and boredom cause a feeling of tiredness. Drowsiness research is in general terms about boredom and monotonous tasks and their effects of person.

Currently the best way to measure the level of drowsiness of a human being is the multiple sleep latency test [16]. The multiple sleep latency test combines the use of sleep diaries, polysomnography done during one night, and a test where the testee goes into a dark, quiet, and temperature-controlled room to lie down for twenty minutes at a time several times during one day. The aim of the multiple sleep latency test is to measure the sleep latency, or the time it takes for the person to fall asleep. The tests are started 90–180 minutes after the testee has woken up, and they are done with two hour intervals. A minimum of four tests is required for a full analysis.

Various signals are recorded during this test. The most interesting ones are the electroencephalogram (EEG), electrooculogram, and electromyogram. This is the trio that is most often used to measure drowsiness in all kinds of experiments. EEG measures the electrical activity of the brain using electrodes placed on the testee's scalp. Electrooculogram records the movement of the eyes and is used to study the deeper stages of sleep [14]. The electromyogram records the electrical activity of skeletal muscles.

EEG measures the electrical activity of the brain. Multiple electrodes are placed on the testee's scalp to detect neural oscillations. Hans Berger discovered the alpha rhythm in 1929 [10] and since then many more rhythms have been found [56]. The main ones are shown in Table 1 and in Fig. 1. The typical amplitude of these waves ranges from 0 to 200 μV.

Rhythm	Frequency (Hz)
Delta	below 3.5
Theta	4-7.5
Alpha	8-13
Beta	14-30
Gamma	above 30

Table 1: EEG rhytms

The alpha rhythm is the classical pattern that emerges when a subject is awake, relaxed, and has their eyes closed. Any kind of movement or visual stimulus as well as concentrating hard distorts the alpha rhythm. There are individual differences, and it is important to start off with a controlled period of relaxation with eyes closed. Niedermayer and da Silva report that the voltage range for alpha rhythm is usually between 0 and 50 μV, and that values of above 100 μV are uncommon in adults.

A subject's level of alertness fluctuates in the process of falling asleep. The criterion for 'being asleep' in the multiple sleep latency test [16] is that more than a half of a thirty second time-frame should consist of sleep, which in turn is defined as either the presence of delta and theta waves, or as an activity in higher frequency with a lower amplitude than what is measured from delta and theta waves [61]. Niedermeyer calls this the alpha dropout [56] and considers it a first sign of drowsiness.

Another sign of early sleep is the presence of slow rolling eye movements [9]. Typically, the frequency of these movements will vary in the range of 0.25Hz to 0.5Hz, and they are usually detected simultaneously with alpha dropout. They show up in EEG as well as in the electrooculogram. Multiple signals help in detecting the source for various artefacts seen in EEG data. Electromyogram detects chewing, swallowing, and the movement of the tongue. In some cases, heartbeat can also show up in the EEG. This rhythm can be filtered out by combining the data with the electrocardiogram data.

The next sign of sleep is the Vertex wave. It is a sharp transient in amplitude similar to the K-Complex, as shown in Fig. 2, but shorter, and with a smaller amplitude change. This is a sign of deep drowsiness [56]. As for the K-Complex, on a four level scale of sleep, it is a sign of stage two. These four levels form the non-rapid eye movement phases of sleep preceding the rapid eye movement stage that is regarded as the deepest sleep. This
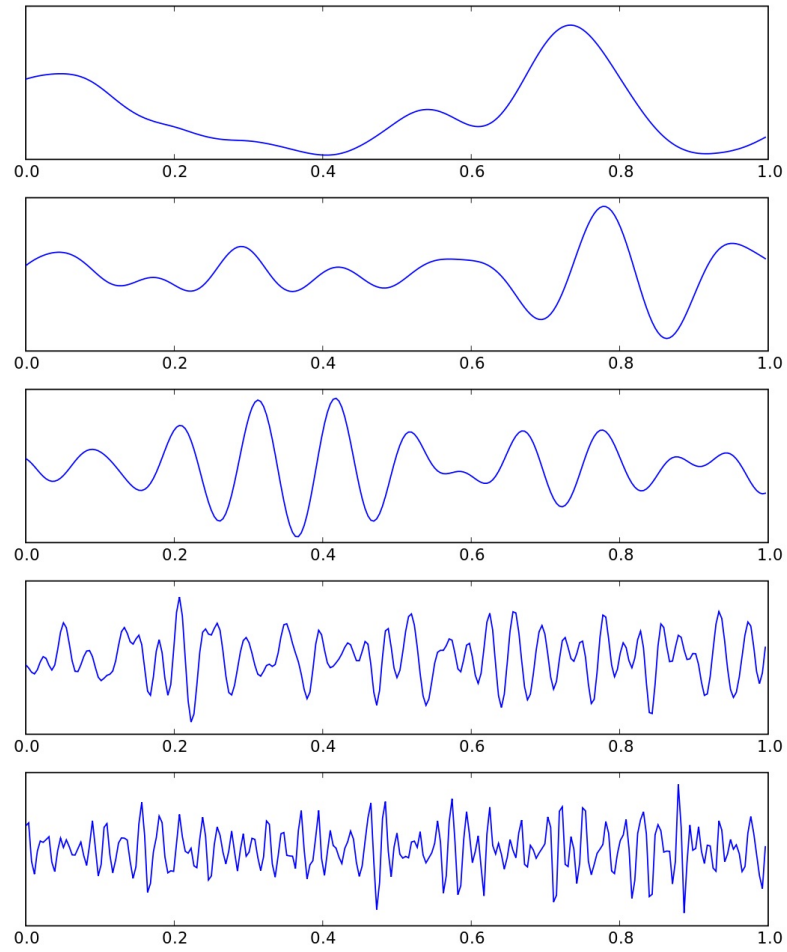
Figure 1: Filtered EEG rhythms. From top to bottom: delta, theta, alpha, beta, and gamma. Copyright © Wikipedia user Hugo Gamboa.
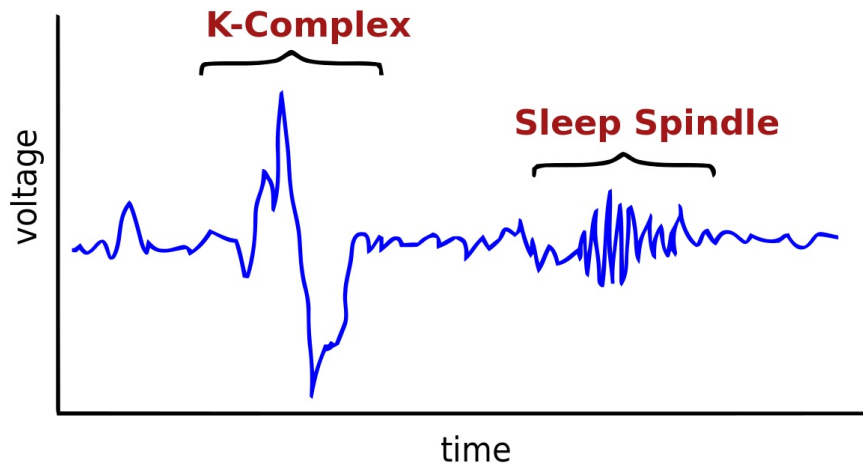
Figure 2: K-Complex and Sleep Spindle are signs of stage two sleep. Copyright © Wikipedia users Ijustam and Neocadre

deepest level of sleep is usually not seen in clinical EEG recordings because it occurs after 100 minutes of sleep [9] and typical trials last for twenty to thirty minutes. In the multiple sleep latency test, each separate test takes twenty minutes.

In addition to vertex waves, the researches look for an oscillatory activity on the range of twelve-to-fourteen Hz lasting at least 0.5 seconds. This phenomena are called the sleep spindle [61] and an example of such is shown in the Fig. 2.

Another sign of sleep is the relaxation of skeletal muscles that is usually seen in stage two. A drop in the amplitude recorded by an electromyogram is a strong indication of sleep resulting in almost total inactivity and paralysis in the rapid eye-movement phase of the sleep. This relaxation effect is shown in Fig. 3 [58]. While the recording in Fig. 3 is from a rodent, the researchers insist that this effect is visible in all mammals [51].

This has been a gross oversimplification of the EEG. For example, the location of the electrodes has been completely overlooked so far. Different signals are measured best at different locations and this allows for vast variations in the number of electrodes used in task specific EEG recordings. The International Federation of Clinical Neurophysiology insists that at least twenty-four channels should be recorded and preferably thirty-two
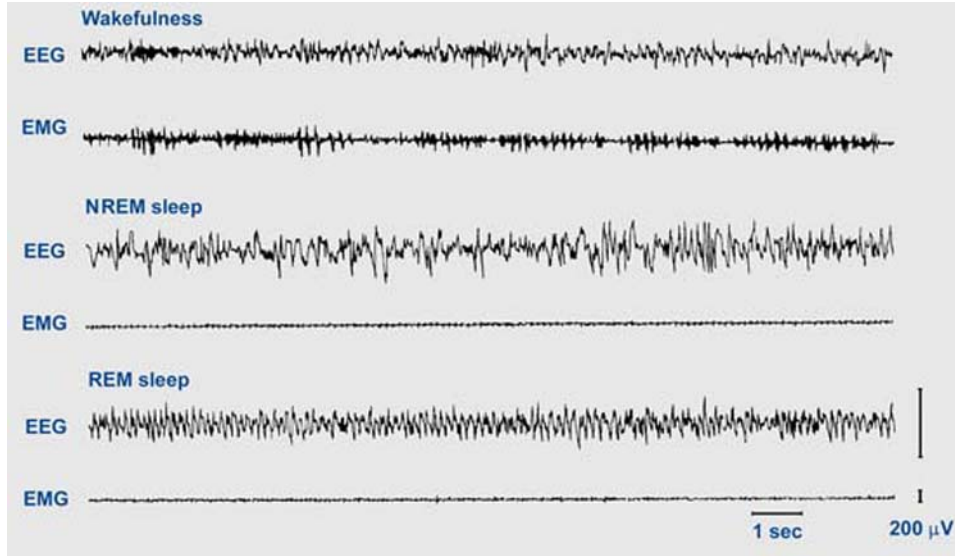
Figure 3: EMG recording of a rodent from Scholarpedia. Copyright © Christopher M Sinton [51]

[57]. Alpha rhythm is posterior dominant and beta more varied, but more frontocentral, so one might not need as many channels to detect the alpha dropout for example.

To recap, this thesis is about detecting drowsiness resulting mainly from lack of sleep. Tiredness falls into this category, while fatigue is somewhat different. The point in this study, however, is not to compare what causes what, but to learn how to detect signs of drowsiness that may be the result of both fatigue as well as of lack of sleep. Many drugs and sleep disorders can also cause drowsiness, but they are not the main focus here [4].

One source of drowsiness is the circadian rhythm. In 1729 Jean de Mairan noticed that plants that were taken to a dark, temperature-controlled room opened and closed just as they do when they are outside and try to gather as much sun as possible [20]. Humans have an endogenously created rhythm that makes us drowsy from two a.m. to five a.m., and from two p.m. to four p.m. This effect persists even when the sleeping pattern is changed to a thirty-six hour day [31].

6

2.2 Biomarkers for Human Drowsiness

Biomarkers are measurable features that can be used to gain knowledge of an underlying physical state. Measuring blood pressure is easier than doing a full electrocardiogram and can be done at home by patients themselves [6]. This allows for faster study and larger sample size when researching problems related to human heart. In addition to electrical activity of the brain and muscles, there are many other ways that have been used to measure drowsiness in a person.

While the focus in the previous chapter was on sleep, here I introduce biomarkers for drowsiness detection. Microsleeps are short unexpected intrusions of sleep [4]. Previously sleep latency was defined as the first thirty second period with more than half of sleep. However, this definition is based on a clinical study of sleep loss. Many studies have focused on finding these short sleep patterns in all kinds of settings.

Where an electrooculogram measures the eye movement, much effort has been made on detecting blinking by using an electromyogram and a video recordings, and even an infrared sensor [15]. Stern, Walrath and Goldstein reported that in a two-hour vigilance task, the number of blinks per minute increased by 43% from the first half an hour to the last [66]. They noted that the rate of blinks depends on the task at hand and reported of a study comparing the rate of blinking in tasks involving mental arithmetic, reading and conversation. In that study, it was found that doing mental arithmetic produced the highest rate of blinks and reading the fewest. Stern, Walrath and Goldstein considered the eye blink a suitable indicator for detecting drowsiness.

Caffier, Erdmann and Ullsperger studied drowsiness from the eyes in a sixty people study where testee's took part in two sessions where they stared at a picture of their choosing for twenty minutes at a time[15]. One session was done before the beginning of the testee's ordinary working day and another after their working day had ended. The assumption was that the testees would be more tired after work. All but three subjects confirmed feeling this way. The mean duration of the testees' eye blinks rose from 202 milliseconds in the morning to 259 milliseconds in the evening. As for the the frequency of blinks, it decreased slightly from 16.33 to 15.84 blinks per minute.

PERCLOS [73] is a measure for how long the eyes are more than eighty

percent closed. This measure detects the increase in frequency or duration of blinks and is particularly interesting for detecting drowsiness in drivers, since a driver's position remains fairly stable for a long time and the camera is able to locate the eyes easily. The problems related to this measuring method concern the lighting conditions and glasses.

A reliable way to detect drowsiness is the psychomotor vigilance task that measures the time it takes for a subject to react to a visual stimulus [22]. Often the stimulus used in tests is just a dot that appears on a screen, and the testee should react to it as quickly as possible by pressing a button. Two metrics are then derived from this. One is the delay between the appearance of the dot and the pressing of the button. Another is the number of lapses. If the person fails to press the button within 500 ms of the stimulus, then it is considered a miss.

The test takes ten minutes. Shorter versions have also been created, but they are not as reliable. Astronauts on board the International Space Station use a three-minute version to get an estimate of their drowsiness [7]. However, Loh, Lamond, Dorrian, Roach and Dawson studied two and five minute versions of the test and found that they are less sensitive than the full ten-minute test. They consider nevertheless, that the five-minute test can be used if the full test can not be conducted [46].

Graw, Kräuchi, Knoblauch, Wirz-Justice and Cajochen did a study of how staying awake affects reaction times with sixteen people, where one group stayed awake for forty hours straight, and another group alternated between a 150-minute period of scheduled wakefulness and a 75-minute period of sleep and repeated this for ten times [32]. The study showed that psychomotor vigilance task performance follows the circadian rhythm and is not dependent on light or on how long the subjects had been awake.

Another method that is often used for measuring drowsiness is self-reporting.

Murray Johns developed a questionnaire for measuring daytime sleepiness [37]. In the questionnaire, the respondents are asked to rate how likely it is that they would doze off or fall asleep in eight different scenarios. The scenarios described to the respondents are: (1) sitting down and reading; (2) watching television; (3) sitting in a public place like at work or in theatre; (4) as a passenger in a car for an hour; (5) lying down in afternoon; (6) sitting and talking to someone; (7) sitting quietly after lunch; and finally (8)

sitting in a car that is stopped in traffic for few minutes. The scale used in the questionnaire, the so-called Epworth sleepiness scale, reaches from zero, 'would never doze', to three, 'high chance of dozing'.

Another scale used for measuring sleepiness by self-reporting is the visual analog scale devised by Timothy Monk [54]. It consists of a straight line on which the testee is to place a mark showing their mood or level of alertness. This is in contrast to a Likert like scale, such as the Epworth sleepiness scale, where the answers are labelled and predetermined. The visual analog scale is divided into a hundred steps with no cues. This allows for greater sensitivity in repeated tests, where the respondent might wish to indicate for example that the level of pain has gone up slightly or that they feel a little bit less alert than before, but not quite enough to move to the next level of the predetermined scale.

The Stanford Sleepiness Scale divides the respondent's sleepiness into seven levels ranging from one for 'wide awake' all the way to seven for 'feeling like falling asleep immediately and having dream like thoughts' [34]. Levels two and three correspond to 'awake' and 'responsive', four to 'somewhat foggy' and five for 'foggy'. Six corresponds to feeling sleepy, but still fighting the sleep. This type of a scale allows for the subjects to be tested almost continuously. In original study the subjects were indeed asked to rate their alertness every fifteen minutes.

The Karolinska sleepiness scale is similar to the Stanford sleepiness scale, but divided into nine levels instead of seven [2]. The meanings of these levels are given in Table 2. In the original study, levels 2, 4, 6 and 8 had only ticks, but no labels. The answers given by the responents were validated against spectral analysis of EEG and visual analysis of electrooculogram in search of slow rolling eye movements.

Another interesting study searching for biomarkers of sleep is the one done by Seugnet, Boero, Gottschalk, Duntley and Shaw [64], where the amylase found in saliva was used to detect increasing levels of sleep debt. In the original study, the levels of amylase were studied on flies which were kept awake with caffeine and methamphetamine. After identifying amylase as a candidate for a biomarker indicating sleep debt, the research group measured the amylase levels of human subjects who were kept awake for twenty-eight hours during a two-weekend test period. Clearly this needs further studying. It is obvious, however, that there is a keen interest for finding a biomarker

that could indicate the level of drowsiness in people in a convincing manner.

2.3 Statistical Methods in Drowsiness Detection

In this chapter, I will show four different studies that apply statistical methods to measured EEG signals in order to detect signs of drowsiness. After these, I will focus in more depth on a study where a group is analysing speech using support vector machines, a machine learning technique and finally a study where a balance board is used to measure the postural balance and this data is then used to measure drowsiness.

Vuckovic, Radivojevic, Chen and Popovic [71] used an artificial neural network to recognise alert and and drowsy sequences. They used forty-two features from a cross-spectral density analysis to train an artificial neural network to detect drowsiness from an EEG signal. First they used two neurologists to find sixty alert and sixty drowsy one-second epochs for each seventeen subjects. Only epochs that were selected by both neurologists were used as a sample. The subjects were allowed to proceed to stage one of sleep, but woken up before advancing to further stages.

Karolinska Sleepiness Scale	
1	extremely alert
2	very alert
3	alert
4	rather alert
5	neither alert nor sleepy
6	some signs of sleepiness
7	sleepy, no effort to stay awake
8	sleepy some effort to stay awake
9	very sleepy, great effort to stay awake, fighting sleep

Table 2: Karolinska Sleepiness Scale

Short periods of sleep were excluded from the material and only samples without signs of sleep were used. The research group performed four experiments with sixty clear samples of alert and drowsy EEG, and finally a five-minute recording of raw EEG signal selected from the thirty-minute recording per each subject. In the first four experiments they used twenty of the selected sixty samples as a training set and forty alert and drowsy as validation set. In a final experiment the artificial neural network was measured against the two experienced neurologists.

Table 1 lists typical Alpha activity in alert subjects as activity in the range of 8-13 Hz. A cross-spectral density analysis [26] groups the activity of each spectrum and thus reveals if there is activity below or above the Alpha

range. The forty-two derived features were then fed into the neural network in order to produce a classifier that would identify if given a one-second epoch shows sign of drowsiness or not.

Three different neural networks were used. Linear network trained by Widrow-Hoff rule, feed-forward network trained with the Levenberg–Marquardt learning rule, and a learning vector quantisation rule were used. In these networks the data flows only to one direction and the difference is how the weights of the connections are updated.

The Widrow-Hoff rule [72] uses a least mean squares approach by starting with zero weights and updating them with the use of the mean square gradient. The Levenberg–Marquardt algorithm solves the problem of minimising a nonlinear function and this helps in finding the optimal nonlinear activation function for each neuron. Learning vector quantisation is a prototype-based supervised classification algorithm that was invented by Teuvo Kohonen [39]. First a self-organising map is used to find prototypes which are then learned by using the learning vector quantisation.

This method requires that the used epochs are selected in such a way that the groups of drowsy and alert are clearly separable. Mardi, Miri Ashtiani and Mikaili [50] has however developed a method that doesn't have this limitation. The researchers calculated fractal dimension of the EEG signal using Higuchi's algorithm [33] and the Petrosian's algorithm [59]. A third feature used was the sum of squared logarithms of energy. These three features were used to train a feed-forward artificial neural network.

The research group recorded three or four forty-five-minute simulated driving sessions with ten drivers, who had been awake for at least twenty hours before the test. The drivers drove in a simulator, trying to avoid obstacles placed along their way. If the driver managed to pass the appearing barriers, they were considered to be in a state of alertness. If crashes happened, the driver was considered drowsy.

The four seconds leading up to a crash were always labelled 'drowsy' and the two following seconds as 'alert', since the simulator played a loud alarm for each crash, thus dispelling the driver's drowsiness. Each crash event was reviewed, and if the subject said that they had been vigilant or if the video showed signs of drowsiness even though the person passed the barrier, then the event was not used as a 'drowsy' example.

Researchers used a band-pass filter to separate different frequencies of

the EEG and extracted two-second epochs that were synchronised to show either a passing or a crash. Each drive was then split into two-second epochs. Finally, altogether 1686 observations were used. 80% of the data was used to train the classifier and the remaining 20% to test the performance of the classifier. This classifier was then used to achieve a two-class classification with an accuracy of 83.3%.

Independent component analysis [38] seems promising tool for finding a signal from a noisy source. The classical example used is the so-called cocktail party problem [17]: being in a noisy room with many people and wanting to hear the voice of the person you are talking with. Similarly, an EEG signal is recorded from many points of the scalp, and independent component analysis can be used to separate the different signals.

Everything shows up in EEG recordings: blinking, movements of the tongue, heartbeat... While driving can seem like a monotonous task, there are all kinds of things that happen inside a person who's driving. This is where the independent component analysis brings out its best qualities. It helps to distinguish local fluctuations, caused by various sources, from the underlining base rhythm.

Lin, Wu, Liang, Chao, Chen and Jung developed an estimator for drowsiness using independent component analysis to filter out noise from measured EEG signals and by linking the brain activity to actual driving performance with the use of a power-spectrum analysis and a linear regression model [44]. They measured EEG from sixteen drivers driving a car for forty-five minutes. The measurements were made in the early afternoon hours, when the circadian rhythm has its second peak.

The researchers created their own metric for estimating driving performance. The driver was instructed to keep the vehicle as close to the centre line as possible, while the simulator was made to cause seemingly random events of sudden changes in steering wheel angle in order to simulate driving on a bad road. These deviations were used to measure the corrective actions that the driver undertook in order to keep the vehicle on course. Five of the sixteen drivers were selected for further study because of the microsleep intrusions detected in their first driving session.

The research group found that drowsiness fluctuates with cycle lengths of more than four minutes and used a moving average filter advancing two seconds at a time to eliminate variance at shorter cycle lengths. After

filtering out the EEG artefacts using independent component analysis the power-spectrum of the EEG was analysed. From the data thus produced, a correlation coefficient was computed using the Pearson product-moment correlation coefficient. The group found out that the log power spectra of two components in the range of 9–25 Hz had a strongly positive correlation with the performance of the driver.

Overall, the group had a success rate of 86.2% for the training session and of 88.2% for the test session when two dominant independent component analysis components were used as input features to a linear regression model for the selected five subjects. The researchers noted that the relationship between the brain activity shown in the EEG data and actual driving performance remained stable from session to session for each driver, but varied somewhat between different drivers.

The biggest obstacle for using EEG for drowsiness detection is the difficulty in getting a signal from the driver. EEG requires twenty-four to thirty-two electrodes and it's not feasible to do a large scale study with the current equipment. All kinds of devices have been proposed, and currently there are more than ten projects on the crowdfunding website Kickstarter seeking funding for an EEG recording device.

Another possible use for EEG signals is the brain-computer interface. Lin, Chang, Lin, Hung, Chao and Wang have proposed a lightweight headband [43] that uses EEG signals to detect drowsiness. The aim of the group was to develop a brain-computer interface that would have the same benefits as devices that focus on the physiological signs of drowsiness, such as eye cameras and grip strength. The main benefit of these is that they don't require contact and are thus easier to accept.

The group also wanted to obtain the qualities of another group of devices that measure the physiological signs of drowsiness, like the electric potential of skin and EEG activities. These devices are less dependent on things such as lighting conditions, vehicle type or driving conditions.

A signal reading module consists of the following components: three EEG electrodes that read brain waves in the occipital midline, a signal amplifier, a band-pass filter, an analogue-to-digital converter and finally, a bluetooth module that sends the processed data to a microprocessor that can then alert the driver in real-time if dangerous drowsiness is detected.

One possible source for signs of drowsiness is a person's speech. Krajewski,

Batliner and Golz studied how speech changes when a person gets more tired and applied machine learning to detect these signs of drowsiness [41]. The group used a support vector machine that was trained to classify speech into alert and drowsy with good accuracy.

The researchers hypothesised a number of changes that they thought would be caused by drowsiness. Their first assumption was that cognitive speech planning would become impaired as a person grows tired, thus affecting the speed of speech as well as slurring the articulation and increasing errors. A second assumption was that the decreased muscle tension that causes flat and slow respiration would cause observable changes in phonation and articulation. Thirdly, the relaxed open mouth display caused by drowsiness would have an effect in the first and second formant position.

Krajewski, Batliner and Golz recorded twelve student volunteers, who participated in an all-night study with an hourly routine consisting of forty minutes of driving, fifteen minutes of vigilance tests, responding to the Karolinska sleepiness scale questionnaire, two minutes of recording speech and a break. For each student, seven sessions in total were conducted between one a.m. to eight a.m. The voice recordings made during the night were split into alert and drowsy samples by combining them with the information provided by the Karolinska sleepiness scale questionnaire, with a threshold of 7.5. This threshold was validated by the presence of microsleeps, which did not occur during alert phases.

Speech data has usually a high number of features because often data is split into ten millisecond frames and smoothed. Data was then analysed frame by frame using signal processing, speaker-recognition and speech-recognition-based acoustic features. The final dataset had over forty-five-thousand dimensions, which were reduced to two-hundred-and-thirty features by using a simple correlation-filter-based feature subset selection.

Finally, the group compared thirteen different classifiers, and the support vector machine managed to correctly classify 86.1% of the samples into alert or drowsy. Class-wise, the average classification performance was 82.8%. Logistic regression [48] was a close second, while the two nearest neighbours classifier [18] and the random forest classifier [13] had the worst performance with overall performance of 62.8% and 68.1% and class-wise averaged performance of 69.5% and 62.9%, respectively. These results required the speaker to be known.

Another promising idea for detecting drowsiness is to use the Nintendo® Wii Fit balance board to measure a person's postural steadiness [68]. This has been tried out in one study, where the balance board was used to measure the postural steadiness of fifteen test subjects every hour during a twenty-four-hour test period. The tested subjects were offered a snack or a light meal every two hours.

The score derived from measuring the subjects' postural steadiness was tested against a 'drowsiness baseline' that the group created by using the Stanford sleepiness scale and an estimation of alertness based on a three-process model that combines homeostatic, circadian and sleep inertia components. The three-process model predicts that alertness peaks seven-and-a-half hours after waking up and at half-past-three in the afternoon, and that drowsiness peaks after staying awake for twenty-two-and-a-half hours and at half-past-six in the evening [1].

The group used multiscale fuzzy sample entropy [76] to extract features from the signal originating from the balance board. The result was a complexity index, which measures the predictability of the signal. Each test consisted of four thirty-second tries during which the person stood on the board and tried to maintain their balance.

For the times considered as alertness peaks by the three-process model, the complexity index group average was 8.9 ± 1.3. For the drowsiness peaks, it was 7.9 ± 1.4. The values given by the Stanford sleepiness scale were 2.0 ± 0.7 (from one to three) for the alert peaks, and 4.8 ± 1.1 (three to six) for the drowsy peaks. Overall, the correlation between the complexity index and predicted alertness was $p = 0.94$.

The most promising thing about using the balance scale for measuring drowsiness is that it is very practical in field use, similarly to the breathalyser used for measuring drunkenness. Law enforcement personnel could carry the board with them, and driving sleepy could become as easily penalised as driving drunk.

Next I will show how the results and methods presented here are used as a baseline in order to study drowsiness in drivers. Many studies take place in a laboratory settings where quick drowsiness detection methods like psychomotor vigilance task and questionnaires precede and succeed periods of driving. EEG and self-reported drowsiness can be recorded when studying real driving.

3 Evaluating the Driving Performance of the Driver

Here I will show how metrics for driving performance have been developed using biomarkers from previous chapter as a baseline. I will first take a look at how drowsiness affects driving performance. I will then describe ways how drowsiness can be detected with measurements made in the car that is being driven. Finally, I will take a look at current research and modern solutions for detecting drivers' drowsiness.

3.1 Psychological Issues Related to Driving Performance

Driving is a complex task that requires situational awareness [25]. A car moving 100.8 km/h travels 28 metres per second. Drowsiness in such speeds can pose a real danger. A project led by the European Commission identified a number of psychological symptoms related to fatigue [63]. They listed such symptoms as decrease in vigilance, slower reaction time, poor memory, worse psychomotor coordination, slower information processing, and poor decision making.

Studies have shown that driving in a monotonous environment causes large steering wheel movements to become more frequent [67]. This is consistent with what the SafetyNet project has also reported. Drivers tend to notice when they get tired and try to compensate this in two ways: either by driving faster to increase the demandingness of the task, or by slowing down and keeping a bigger distance to other vehicles in order to compensate for longer reaction times [63].

The circadian rhythm also affects the driver [42]. Due to this evolution-based rhythm, a human's desire to sleep is strongest from 2 a.m. to 4 a.m. and again between 2 p.m. and 3 p.m. Alertness is highest at 10 a.m., eye-hand coordination at its best at 2:30 p.m. and reaction times shortest at 3:30. These aspects are further discussed below.

Wright, Hull and Czeisler studied the relationship between alertness, performance, and body temperature [75]. In their study, the testees' day was stretched to a twenty-eight-hour cycle consisting of eighteen hours and forty minutes of scheduled wakefulness and nine hours and twenty minutes of scheduled sleep. The fourteen participants were made to live by this rhythm for twelve cycles in total, corresponding to fourteen normal days. The subjects were asked to do a thirty-minute battery of tests every two
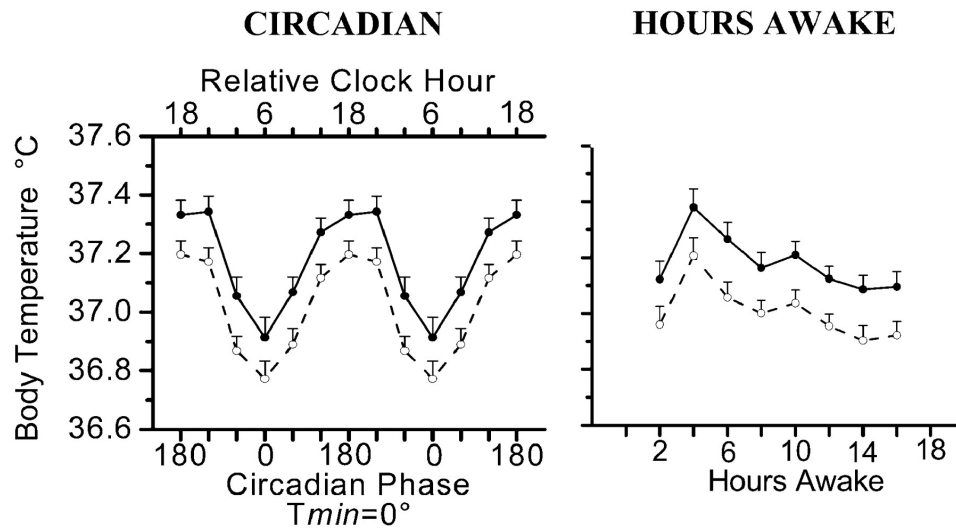
16

CIRCADIAN   **HOURS AWAKE**

Figure 4: Average high and low body temperature of 14 subjects across 24 h circadian phase (left) and hours awake (right). Solid line for High and double dotted for low. Error bars represent ± standard error. Copyright © 2002 the American Physiological Society

hours, starting two hours after the beginning of the scheduled wakefulness.

The researchers knew that the human body cannot maintain a twenty-eight-hour day under selected conditions, and that it will continue its twenty-four-hour cycle thus enabling the researchers to study the circadian rhythm with different waking times. The participants body temperature, which was measured once a minute, varied roughly by one degree celcius, while the temperature and lighting conditions of their surroundings were kept unchanged. Sixty degrees is four hours in this twenty-four-hour clock face.

Dijk, Duffy and Czeisler studied alertness with two groups of participants. One group was kept awake by a constant routine for a total of thirty-six to sixty hours after an initial synchronisation phase. This group showed fairly consistent mental performance for the first sixteen hours. After that, their mental performance correlated with the circadian rhythm and with the low point of the temperature cycle [21].

The other group had a twenty-eight-hour routine similar to the one described above. The self-reported minimum level of alertness for this group coincided with the body temperature minimum, and mental performance minimum followed shortly after. The testees' did not show the expected
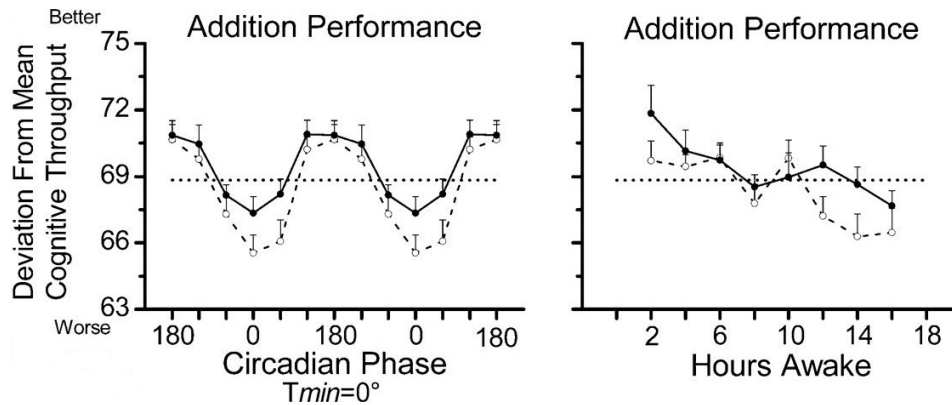
17

Figure 5: Circadian phase-dependent (left; data double plotted) and hours awake-dependent variation (right) of addition performance. Neurobehavioral data are expressed in deviation from individual subject's mean. Scores in the upward direction represent better performance. The group mean (n = 14) is added to the high-low deviation scores to indicate the amount of change in performance. Error bars represent ± standard error. Dotted line represents the group mean. Copyright © 2002 the American Physiological Society

mid-afternoon drop in either alertness or mental performance. The testees' mental performance was assessed by having them perform an hourly test that consisted of adding pairs of two-digit numbers.

Wright, Hull and Czeisler had their participants do this same task as well, and Fig. 5 shows how the performance of their testees varied according to different phases of the circadian cycle and to hours of staying awake [75].

Being able to react fast in surprising situations is vital for drivers. The fact that reaction times vary according to the time of the day is also essential to keep in mind while searching for ways to improve driving safety. Reinberg, Bicakova-Rocher, Nouguier, Gorceix, Mechkouri, Touitou and Ashkenazi [62] studied the daily variation of reaction times by comparing those of the driver's dominant and non-dominant hand. Their hypothesis was that there is a difference in the rhythm of the daily variation of reaction times between these two hands, and their research showed that this is indeed the case.

The group tested eleven testees with a simple reaction test: a yellow light was shown to the testee, and the time it took for the testee to press a button after seeing the light was measured. A three-colour setup was also used, where the subject was required to press the button with their left hand at the appearance of one colour, to press it with their right hand at the
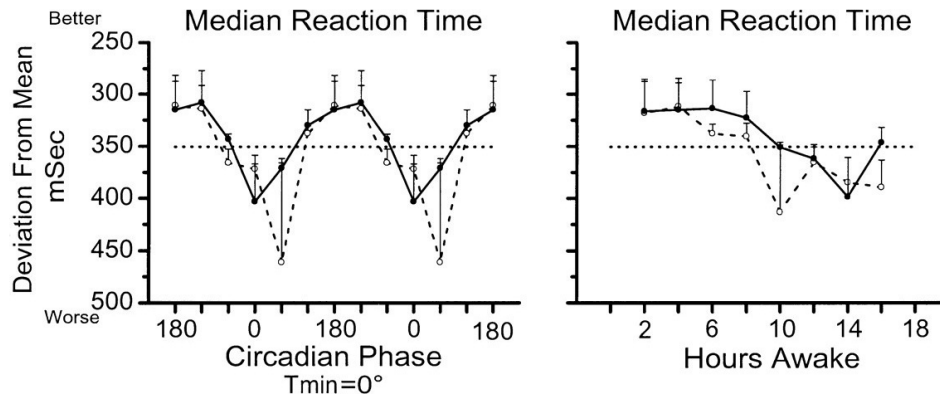
Figure 6: Circadian phase-dependent (left; data double plotted) and hours awake-dependent variation (right) of median reaction time in Psychomotor Vigilance Task performance associated with high vs. low body temperature. Copyright © 2002 the American Physiological Society

appearance of a second colour, and to ignore the third colour. The test, in which ninety-six signals altogether were given, was done from four to eleven times a day for each testee, during twelve to fifteen days. The testees' reaction times were noted to vary by around 20% depending on the phase of their circadian rhythm.

Wright, Hull and Czeisler et al. used the psychomotor vigilance task as a part of their assessment for drowsiness and for the effects of the circadian rhythm [75]. Figure 6 shows how the median reaction time was noted to vary across the circadian phase and how this was further influenced by how long the testee had been staying awake. The median reaction time varies from 300 to 400 milliseconds. During this difference of 100 milliseconds, a car driving a hundred kilometres per hour advances three metres.

Figure 7 shows how the number of lapses, or responses that took more than 500 milliseconds to react to a stimulus, increases first, dips during the afternoon hours, then rises again and finally decreases quite rapidly after having been awake for twelve hours.

Edwards, Waterhouse and Reilly studied the effects of the circadian rhythm on alertness with a game of eye-hand coordination. In the game that they devised, a small disc is flicked in order to hit a target at the distance of thirty or fifty centimetres [23].

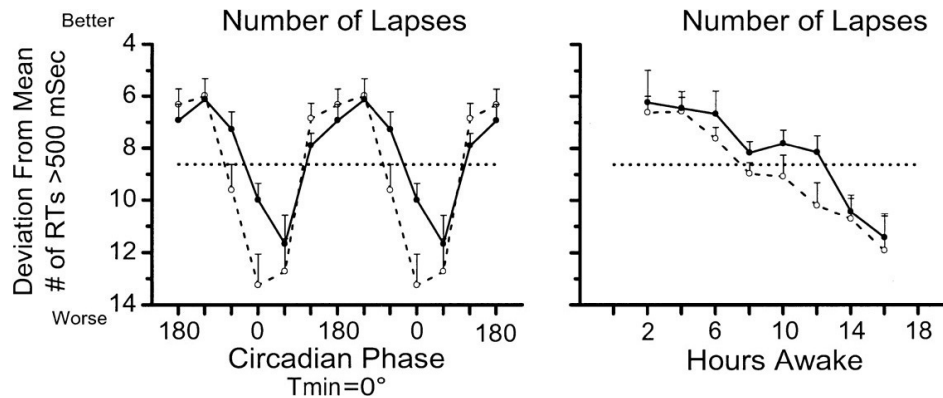In each test session, the seventy students that were tested had twenty

Figure 7: Circadian phase-dependent (left; data double plotted) and hours awake-dependent variation (right) of number of lapses in Psychomotor Vigilance Task performance associated with high vs. low body temperature. Copyright © 2002 the American Physiological Society

attempts at the shorter distance target first, and then another round of twenty tries at the longer distance target. In total, the test took less than ten minutes to complete, and it was performed six times with each student, with 28-hour intervals. Each throw was scored on a scale of zero to six, and the first three and final three throws were excluded to eliminate the effects of a possible anxiety related to starting and finishing the task.

The accuracy of performance for both the short- and the long-distance shot motor task showed a correspondence between the circadian rhythm and the subjects' measured core temperature and self-reported alertness and fatigue scores, which were measured using a visual analogue scale. Performance was at its best at 4:00 p.m., when the range of mean scores reached from seventeen to twenty-four in the short-distance task and from fourteen to twenty in the long-distance task.

The study was repeated to see if there was a difference between the dominant and non-dominant hand in the effect of the circadian rhythm, as suggested by Reinberg et al. A similar pattern was found in this second study, where seventy-eight students were made to perform the same task as described above [24].
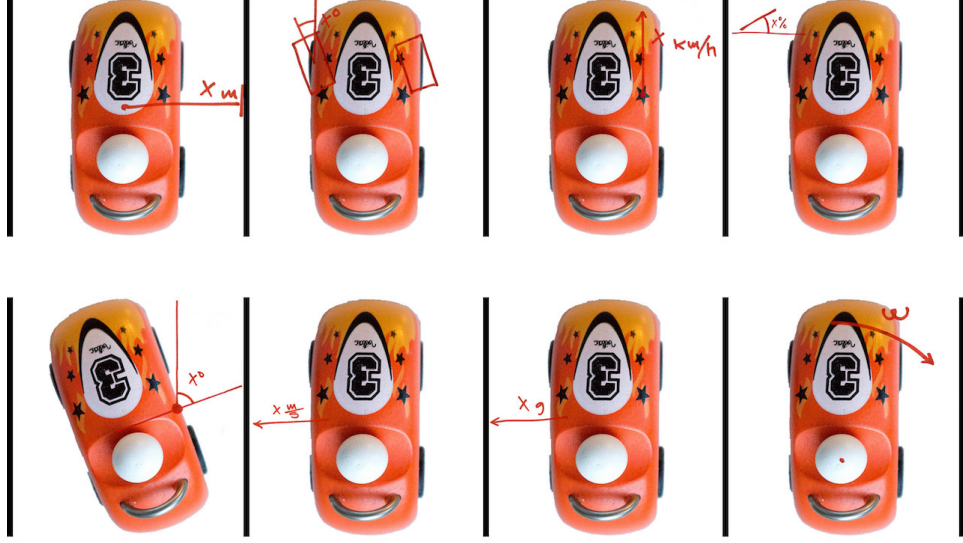
Figure 8: Eight sources of data: (from top left) lateral lane position, steering wheel angle, speed, accelerator pedal position, yaw angle, lateral velocity, lateral acceleration, and angular velocity.

3.2 Biomarkers for Driver Drowsiness

Forsman, Vila, Short, Mott and Van Dongen evaluated eighty-seven different metrics for detecting a driver's drowsiness on the basis of data provided by measurements done on the car that is being driven. The group identified eight different sources for such data [28]: the car's (1) lateral lane position, (2) lateral velocity, (3) lateral acceleration and (4) angular velocity, the car's (5) speed and (6) yaw angle, (7) the position of the accelerator pedal and (8) the angle of the steering wheel.

The lateral lane position is typically measured with a camera that records a view of the road on video. The distance of the car to the lane markings is calculated from this video. The study group listed twelve variables for using lateral lane position to detect drowsiness. The (1) mean value can reveal if the driver starts to compensate for drowsiness by driving closer to the centre line or nearer to the other side. Standard deviation (2) and (3) variance will show how the values are spread around the mean value. Root-mean-square (4) of the lateral lane position can also be used, as well as the (5) signal-to-noise ratio, which here is the reciprocal coefficient of

21

variation of signal variability. These five variables are calculated from all sources of data, with the exception of the yaw angle.

The rest of the variables that can be calculated from the lateral lane position include the three variables that are used to measure signal deviations: (6) the total signal deviation from windowed signal average, (7) the average signal deviation from windowed signal average and (8) the variance of signal deviations from windowed signal average. These three use a twenty second window to detect anomalies in the signal.

Signs of drowsiness can also be observed from (9) the most common signal frequency, as well as from (10) the energy or power of the signal. Finally, there are two variables that estimate how long it would take for the car to cross the lane boundary if no steering corrections are made. These are (11) time to lane crossing within six seconds and (12) time to lane crossing within sixty seconds. The standard deviation of the lane position was stated by the group as the most important indicator for a driver's drowsiness measured from the vehicle, along with steering wheel movements [45].

Among the eight sources of data identified by Forsman et al., the steering wheel angle is the one with the greatest number of variables. All the variables that have been used to detect drowsiness by measuring the lateral lane position have also been used in relation to the steering wheel angle. In addition to these, there are twenty-six variables more that Forsman et al. identified in their review.

Fifteen of these twenty-six variables are related to signal amplitude. Five of these fifteen calculate a percentage of samples with signal amplitude exceeding levels of 0.3, 3, and 30 degrees and 1, and 2 standard deviations. There are variables for maximum signal amplitude to the right and left, cumulative signal amplitudes to right and left, and percentages for samples with signal amplitude to right/left.

Finally there are variables with most common signal amplitude to right and to left and corresponding variables for percentage of samples with those amplitudes. Eleven remaining variables for steering wheel angle are: degree of interaction between steering wheel angle and yaw/lateral acceleration, number of steering wheel direction reversals, average radius in phase-plot, arc length of phase-plot, area of phase-plot, entropy of central/right/left steering movements, fractal dimension of signal, and reaction time from signal auto-/cross-correlation with yaw of the car.

Vehicle speed can also reveal drowsiness. The SafetyNet project discovered that drivers tend to compensate their tiredness by either driving faster to increase the demandingness of the task or by slowing down and keeping a bigger distance to other vehicles in order to compensate for longer reaction times [63]. Mean value, variance and standard deviation have been used to study drowsiness on the basis of vehicle speed, as well as root-mean-square and signal-to-noise ratio.

The time to lane crossing is calculated from vehicle speed for both six and sixty seconds. Fractal dimension shows the complexity of the curve drawn from the values. Basically, this shows if the driver suddenly changes from making many small corrections to making fewer but larger corrections. The hypothesis is that this tends to happen when the driver starts to lose their focus and coordination.

The variables unique to this biomarker are the three measures of deviations from 55 mph. These are (1) an average of signal deviations, (2) standard deviation of signal deviations, and (3) reciprocal coefficient of variation, or signal-to-noise ratio of signal deviations from 55 mph.

The accelerator pedal position is closely related to the speed of the vehicle, but it can show the effects of growing tiredness faster. Mean value, variance and standard deviation have been used to detect drowsiness on the basis of accelerator pedal position, as well as the root-mean-square and signal-to-noise ratio. These five variables have also been used with lateral velocity and lateral acceleration. The latter of these has also been used in connection with the steering wheel angle to calculate the degree of interaction between different types of signals. This degree of interaction is the distance of normalised variables, and it measures how well the driver and the car interact [47].

Angular velocity is a measurement for how fast the car is changing its direction. In addition to mean value, variance, standard deviation, root-mean-square and signal-to-noise ratio, maximum amplitude to right and to left have been used as a possible biomarker for drowsiness. The most common amplitude to both directions, as well as the percentage of values matching that value, have been used in research as indicators for drowsiness.

The phase-plot or phase portrait of the angular velocity values is a source of data used in relation with the remaining three variables identified by Forsman et al. Average radius, arc length and area of the plot can provide

insight and reveal, for example, how steady the velocity of the vehicle is. If the arc length grows, but the mean stays the same, this indicates that there are a lot of changes in the signal value.

The yaw angle, or the direction where the car is heading, measures the orientation of the car in relation to the road. Six variables have been used in previous research to detect drowsy driving on the basis of the yaw angle. The most common signal frequency of the yaw angle can offer better sensitivity compared to the lateral lane position by telling which way the car is headed. The energy of the signal is the root-mean-square to the second power. Much like with variance and standard deviation, the reason for using a multitude of variables has to do with the scale of the values.

The time to lane crossing can be calculated from the yaw angle and vehicle speed as well. This measurement has been used with values of six and sixty seconds. Reaction time can be estimated by cross-correlating it with the steering wheel angle. Finally, the degree of interaction with the steering wheel angle tells how well the driver and the car are in sync. Large corrections that result in lots of smaller corrections for the initial correction can reveal drowsiness.

The aim of the research done by Forsman et al. was to compare different proposed metrics for detecting a driver's drowsiness on the basis of data provided by measurements made on the car, and to develop a method for detecting moderate levels of drowsiness instead of a grave and imminent problem like having fallen asleep at the wheel. In Forsman's study, forty-one tested participants lived inside the laboratory for fourteen days with either a night-shift or a day-shift work.

Participants were tested four times each day. During each testing period, they were first tested with a ten-minute psychomotor vigilance test, were then made to drive for thirty minutes, and another psychomotor vigilance test after the drive. The test was completed by a short neurobehavioral test battery that included a self-reported drowsiness score and a visual analogue scale for the mood along with many different tests that measure for instance reaction times and the ability to substitute digits with symbols in order to detect their fatigue.

Principal component analysis was used to reduce the dimensionality of the data across all participants to include only measures concerning steering wheel variability and lateral lane position. Lateral lane position also correlated the

most with the participants' performance in the psychomotor vigilance test which is an oft-used method for measuring drowsiness. Eight curved road segments were used to validate these findings in each driving session.

They also showed that the lateral lane position could be derived from steering wheel angle measurements using transfer functions. This was an important step because the lateral lane position requires the lane markings to be visible to the camera. The visibility of the markings could become a problem for example on roads covered with snow. Forsman's research suggests that the steering wheel could be used alone to develop a biomarker for driver drowsiness.

3.3 Existing Solutions for Detecting Drowsiness in Drivers

Morris, Pilcher and Switzer proposed a model for detecting drowsy driving with the use of retrospective analysis around curves [55]. According to their research, the yaw angle (vehicle heading difference) performed best in detecting drowsiness that was measured using psychomotor vigilance test and fitness impairment tester. Fitness impairment test measures the eye saccade velocity in a thirty second test. Researchers were able to detect significant differences in driving performance in eight of the ten test sessions.

In Morris' group's study, twenty college students were kept awake for twenty-six hours and made to drive five twenty-minute test sessions in a simulator during one day, starting at eight p.m. and finishing at ten a.m. The testees were woken up at a pre-determined time between eight a.m. and ten a.m., and waking up was confirmed with a phone call. A total of thirty-seven different variables were recorded from the car during the test sessions at a rate of one Hz.

A psychomotor vigilance test was administered on the testees once during each test session, and a fitness impairment test was completed twice before and twice after each session, with a five-minute break in between. Measured reaction times and numbers of lapses were noted to increase in each measurement, as was to be expected on the basis of current understanding on how staying awake affects humans. Saccadic movement of the eye is normal, jerky, motion of the eyes. The speed of the movement slows down as the person gets more tired. The speed of this movement decreased in the testees, as was expected.

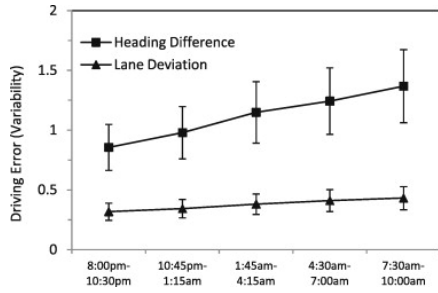The research group measured two things on the car. The vehicle heading

Figure 9: Driving variability over the night for lane deviation and heading difference (mean ± standard error). (Image from Morris, Pilcher, and Switzer [55] © 2015 Elsevier Ltd.)
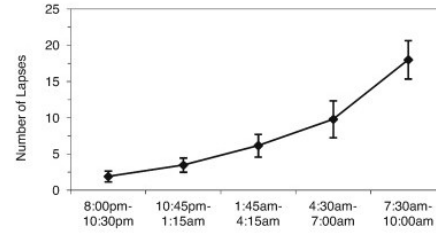
Figure 10: Number of lapses (response time >500 ms) from psychomotor vigilance task (mean ± standard error). (Image from Morris, Pilcher, and Switzer [55] © 2015 Elsevier Ltd.)

difference was noted to be a better indicator than lane deviation (see Fig. 9 above). The research indicated that once people grow more tired they start to make fever but bigger corrections as they drive through turns. The method that was noted to be most reliable was able to detect a significant difference in the means of eight of the ten compared test sessions.

The research group tried both root-mean-square error from the absolute difference and the standard deviation formula for both measurements and the results varied from six of ten with root-mean-square and lane position to the eight of ten with root-mean-square of heading difference. Both got seven of ten with standard deviation formula.

The mean of heading difference seems to grow, but the error is pretty high compared to lane deviation. Figure 10 shows how the number of lapses in a ten-minute test grows from roughly two to more than seventeen on average. Heading difference, measured in degrees, shows the relation of the yaw of the car and the tangential direction of the lane.

In another study, lane departures in a simulator were used as biomarkers and McDonald, Lee, Schwarz, and Brown [53] wanted to learn to classify the last sixty seconds of driving before the event and a matched sample of alert driving on the same road segment. Researchers used supervised machine learning and trained the classifier with carefully chosen samples of driving. Their goal was to develop a method to alert the driver before a lane departure.

26

Seventy-two participants took part in the study. They drove three thirty-minute rides in a simulator. The faces of the testees were filmed as they were driving, and each lane departure was manually scored by using an observer rating of drowsiness [73]. The threshold for a departure to be considered drowsy was thirty-seven, corresponding to 'at least a moderate level of drowsiness'. Each drowsy departure was matched with data from the same road segment, and if possible, from the same driver. The final six seconds leading up to the lane departure were removed because the researchers' aim was to create a system that could alert the driver beforehand.

The drivers' level of alertness was confirmed using a modified psychomotor vigilance task, the Stanford sleepiness scale and a retrospective sleepiness scale. Only drivers who passed as alert in all three measures were qualified as alert samples. The final evaluation data set included eighty alert samples of a same driver and another 336 samples of other drivers on a same road segment as the 162 drowsy sample.

The study group used seven different classifiers and measured the success of each classifier using a receiver operating characteristic, which is a curve of true positive rate in the y-axis against the false positive rate in the x-axis. Two such plots were compared by calculating the area under the curve. Figure 11 shows the random Forest classifier compared against other six classifiers.

Decision tree [35] is used to classify new unseen data based on training data set that is labeled. From the training set a binary tree is constructed by creating splits like 'steering wheel angle is greater than five degrees' or 'distance from the centre point is greater than one metre'. Each split-line is a binary choice and each new data point is then taken to the right leaf where a majority vote decides the label for this data. In this study the depth of the tree was limited to six.

Random forest [13] is a collection of decision trees that have been trained on a different parts of the training set in order to reduce variance. Bootstrap aggregating, or bagging [11], generates new training sets by sampling from the original training set uniformly with replacement. This is called bootstrapping and it's a useful way to estimate the accuracy of an estimator. Here McDonald, Lee, Schwarz, and Brown used five-hundred trees with forty-one features for each iteration.

Another ensemble method is a boosted tree [12]. In this method, the
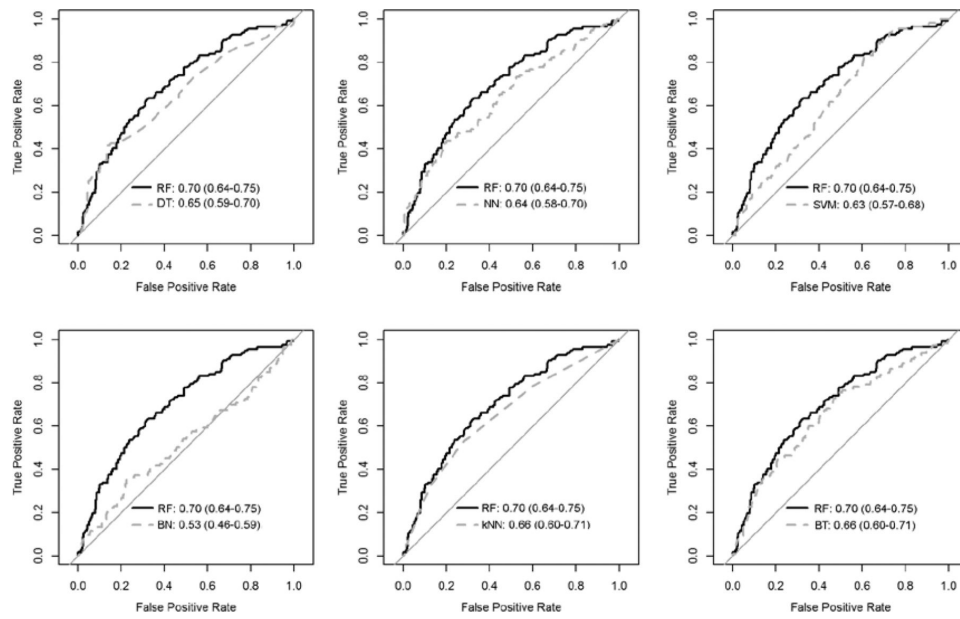
Figure 11: Receiver operating characteristic comparison plots for the random forest (RF) model and the six other models. From left to right, top to bottom, the figure shows the following models: decision tree (DT), neural network (NN), support vector machine (SVM), naive Bayesian (BN), k-nearest neighbour (kNN), and boosted tree (BT). Copyright © 2013, Human Factors and Ergonomics Society

decision tree is trained with training data, but all of the misclassified data is taken away and used to train another tree. This is repeated until all of the training data has a tree that classifies it. For classification, a weighted vote is used to give the new data point a label. Earlier trees are given larger weights, and each subsequent tree gets a lower share of the vote. In this study, the researchers used 0.1 as the decay value.

A fourth classifier tested was a neural network [52]. It is made of interconnected layers of nodes, and each connection has a weight. A node can be thought as a representation of a sigmoid function to the sum of the output of previous nodes. Weights are updated in a process called backward propagation of errors, or backpropagation. McDonald, Lee, Schwarz and Brown used two hidden layers and 0.1 as the training weight decay.

The k-nearest neighbours classifier [18] memorises the training data. For new data, it just takes k closest matches and assigns the label for the data by taking a majority vote on those matches. The search group used nine as the value of neighbours to classify unseen data into alert and drowsy drives.

Naive Bayes' classifier builds upon Bayes' theorem [8] for events A and B. If we detect B and want to know what the probability for A given B is then we can use the Bayes' theorem:

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

to come up with conditional probabilities. The classifier gives the label from the event associated with the highest conditional probability as its answer.

Finally, a support vector machine [69] was used to try and fit a hyperplane between the alert and drowsy samples. This was done in order to produce a classifier that could be used to predict new data by looking at its features. Original data is often not separable, but it can be made separable by adding dimensions which help to classify the data.

In the end, the random forest performed best with an area under curve of 0.7. The other classifiers were in the range of 0.63 to 0.66 with the exception of Naive Bayes', which performed worst. The Naive Bayes' classifier was the only one that was beaten by the random forest with statistical significance.

The random forest classifier was then compared to an algorithm that measured when the eyes of the driver were closed. It had a three-minute window and a threshold of seventy percent for eye closure. It calculated how

much of the three-minutes the eyelids covered more than seventy percent of the eyes. This setup gave the best area under curve of 0.63. Mean accuracy was 0.55 and 0.79 for the eye closure algorithm and random forest classifier, respectively.

A lot of research has focused on detecting drowsiness from EEG data. Yeo. Li, Shen and Wilder-Smith presented an EEG-based drowsiness detector [77] that reached 99.3% classification accuracy for unseen data, with human scored drowsiness used as a benchmark. In Yeo's study, twenty students from the National University of Singapore drove in a simulator and had their EEG recorded with the use of a thirty-two-channel system with 256 Hz sampling frequency. The one-hour drive was done during the afternoon sleepiness peak at three p.m.

The research group used support vector machines to learn how to detect drowsiness from EEG data after carefully selecting the training data by looking at eye blinks and EEG frequency. They then extracted four variables: dominant frequency, average power of dominant peak, centre of gravity frequency, and variability. Finally, they used two different raters and only included epochs where both of these agreed on the alertness or drowsiness of the epochs.

4 Learning to Predict Sleepiness on the Basis of Car Data

In 2013, there were 197 motor vehicle accidents in Finland that resulted in loss of life. Thirty-nine passengers and 173 drivers were killed in these accidents [40]. The Finnish Motor Insurers' Centre publishes statistics on the causes of motor vehicle accidents, and according to them, seventeen of the accidents that happened in 2013 were caused by loss of alertness or because the driver had fallen asleep [40]. According to the estimates of the World Health Organization, there are 1.25 million deaths worldwide each year that are caused by traffic accidents [74].

In Finland, it is illegal to drive if you are too tired. However 92.5% of the drivers punished under Article 63 of the Road traffic act in 2004 and 2005 were involved in an accident [60]. Radun, Radun and Ohisalo found no cases where the driver would have received a penalty in a random stop. There appears to be no way to detect dangerous drowsiness before the driver

actually falls asleep.

Professional drivers transporting commercial goods or passengers use tachographs to record their driving hours, which are limited by European Union regulation. This is currently the best available tool to prevent drowsy driving. People continue driving even when they know they are too tired to drive. In a 1520-person study, conducted by TNS Gallup in the spring of 2013 for the Finnish Road Safety Council Liikenneturva, 26% of the drivers reported having fallen asleep at least once and 95% having driven even when they knew they were tired [49].

Earlier on in this work, I described the research led by Pia Forsman [28] in which eighty-seven sleepiness indicators were identified that had already been observed in various studies over the years. I will now show how supervised machine learning can help in detecting drivers' drowsiness.

Previous research has mainly focused on detecting dangerous drowsiness just before an accident. As seen in previous chapters, all kinds of data has been used for this purpose, produced by measurements performed both on the driver as well as on the car. Here my aim is to use data recorded on a daytime drive to predict which drivers will fail to continue their driving because of dangerous drowsiness. Data about the driver will only be used to train the classifier, while for the actual prediction, the classifier will be shown data concerning the car, or more specifically, the car's steering wheel, speedometer and accelerator pedal.

4.1 Data Source and Measurements

The data used in this study is from a study done by Torbjörn Åkerstedt, David Hallvig, Anna Anund, Carina Fors, Johanna Schwarz and Göran Kecklund, who used the same data in their study on driving at night and dangerous sleepiness [3]. They studied the development of indicators for dangerous drowsiness that led to the termination of a night-time drive by either the drivers themselves or by the accompanying test leader .

The group did their study with eighteen individuals who were randomly chosen among the vehicle owners registered in Sweden and living in the region of Linköping. These testees were first made to drive during the afternoon for roughly eighty kilometres towards Nyköping, and then back along the E4 motorway. After seven hours of resting in a controlled environment, they were made to drive the same route again. In the original study, EEG and

electrooculogram were identified as the best sources of information about drowsiness.

Each driver had a test leader alongside them when they drove. The setting was fairly strict: any request for a break resulted in the termination of the drive, and the driver was instructed not to speak to the test leader. The only moment when the driver was allowed to talk was when the test leader asked them to evaluate their feelings of drowsiness with a number from one to nine, which was done every five minutes.

Eight of the eighteen night-time drives were terminated prematurely – two by drivers themselves and six by the test leader. Three of these eight reported having fallen asleep and, alarmingly, one of the ten drivers who finished the drive also reported as having fallen asleep during the drive even though the test leader had had a view of the driver's face. There was no sudden increase in any of the indicators immediately before the critical event, but rather a steady increase in several indicators.

The daytime drive was done between 15:30 and 19:15 and the night-time drive on the following night, between 00:15 and 04:15. The wide window for the sessions to take place was caused by the fact that two drivers were tested each day. All of the drivers were also asked to keep a sleep diary during the three days preceding the day of the driving sessions.
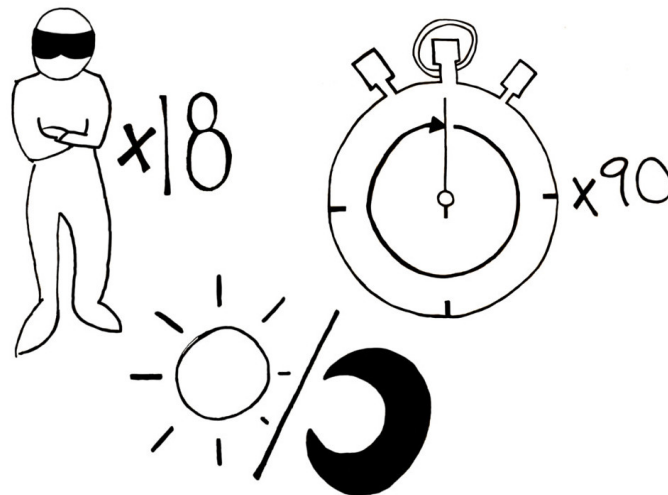

Figure 12: Experiment set up showing eighteen drivers driving for ninety minutes at a time first during the day and then during the following night.

The aim of the study was to detect moderate levels of drowsiness instead of the extreme dangerous ones. The drivers were not allowed to drink caffeine or to listen to the radio, and if they wanted to take a break, the drive was terminated. The test leader travelled along during all of the rides, and the car was equipped with two sets of controls in order to allow the test leader to take over should the driver fall asleep. After the first drive, the testees were served dinner and then taken to a controlled environment where they could watch the television, read and interact socially with the research group.

Two sets of measurements were made during the drives. First, a number of measurements were made directly on the driver. These included EEG, electrooculography, electromyography and electrocardiography. The sampling frequency was 512 Hz for EOG and 256 Hz for others. Another set of measurements were made on the car that was being driven. Ten variables were produced by recording data from the GPS signal, while twenty other measurements related to the direction of the car and its position on the road, such as the distance travelled or the direction of the yaw of the car. The sampling frequency for measurements done on the car was 10 Hz. Table 3 lists all the variables produced with these measurements.

The researchers derived two additional variables from the measurements made on the driver. They calculated the Karolinska Drowsiness Score (KDS) [30] and used an algorithm developed by Jammes, Sharabty and Esteve [36] to calculate blink durations. Each twenty second epoch of driving was manually scored for signs of drowsiness in either the driver's brainwaves or in the electrooculogram. Slow, rolling eye movements of at least $100\mu V$ and lasting for more than one second were marked as signs of drowsiness in the electrooculogram.

Each epoch was split into ten units of two seconds, after which the scorer searched each two-second segment for signs of drowsiness. The KDS for the whole twenty-second epoch is the number of these smaller segments that show physical signs of drowsiness so that an epoch with two segments containing drowsiness events gets a KDS of 20%.

KAROLINSKA DROWSINESS SCORE
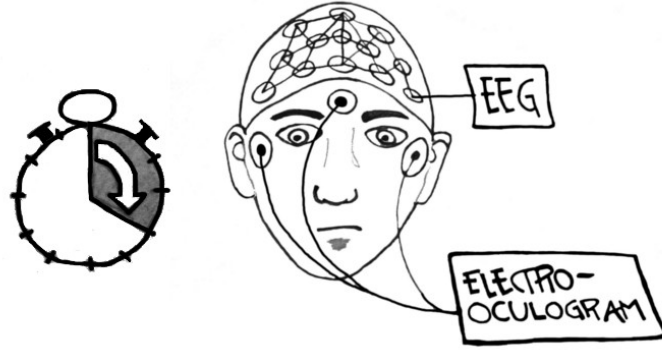
Figure 13: Each twenty seconds is labelled with how much sleepiness the EEG and the electrooculogram show.

Table 3: Controller area network variables based on measurements made on the Car

variable name	description	precision
mAccelPedalPos	How much the accelerator is pressed	0.1%
mAmbientLightCond	ambient light on	boolean
mAmbientTemp	ambient temperature	0.01 °C
mDirInd	direction indication	
	with oscillations removed	$\{0,1,2\}$
mDistanceTraveled	distance traveled	1m
mEngineSpeed	engine speed	0.0001rpm
mGearSelected	selected gear	$\{0,1,2,3\}$
mInCarTemp	in-car temperature	0.1 °C
mLateralAcc	values from (-3.324 to 3.551)	0.0001m/s^2
mLateralJerk	lateral jerk	0.01 m/s^3
mLeftLaneOffset	left lane offset	0.01m
mLeftLaneQuality	left lane quality	boolean
mLongAccBackUp	longitudinal acceleration	0.001m/s^2
mLongAccSensor	longitudinal acceleration sensor value	0.0001m/s^2
mNumFixSatellites_GPS	number of GPS satellite connections	count
		Continued on next page

Table 3 – continued from previous page

variable name	description	precision
mOdometer	odometer of the car	1km
mRightLaneOffset	right lane offset	0.01m
mRightLaneQuality	right lane quality	boolean
mSteeringAngle	steering wheel angle	0.01°
mVehicleSpeed	vehicle speed	0.01km/h
mVideo0Indices	frame index of video 0	count
mVideo1Indices	frame index of video 1	count
mVideo2Indices	frame index of video 2	count
mVideo3Indices	frame index of video 3	count
mWiperFront	front windshield wiper in use	boolean
mYawRate	yaw rate	°/s
timeindex	time index with a step of 100	0.001s
timeSync	time synchronisation index with a step of 100	0.001s
timeNTP	network time protocol time	0.00001s
sleepinessExperiment	marker to when the experiment is on	boolean
speedLimit90	speed limit is 90 km/h	boolean
dmLeftLineCrossings	left line is crossed	boolean
dmRightLineCrossings	right lane is crossed	boolean
dmLane	1 = right lane, 2 = left lane, 12 = switching from left to right, 21 = switching from right to left	{1,2,12,21}

Table 3: Controller area network variables based on measurements made on the Car

4.2 Summary Statistics Used to Train a Classifier

Measured data is split into twenty second epochs and thirteen summary statistics is calculated from these epochs. From the fifty-two available measurements, I chose three as a basis for creating summary statistics. The selected measurements were the angle of the steering wheel, the position of the accelerator pedal and the speed of the vehicle. These were all measured at a frequency of ten Hz, so each twenty second epoch contains 200 values

for each variable.

First the two data sources had to be synchronised. For this, I used the time synchronisation variable in a controller area network and European Data Format data sets. The controller area network contains all the measurements that were made on the car, while the European Data Format variable contains data that was either directly measured on the driver or derived from that data. Most importantly, the European Data Format variable contains the derived KDS measure. All of my summary statistics are based on the controller area network data set.

The zero point in time synchronisation variable marks the moment when the car passed the starting point of the experiment, which was a traffic sign stating the speed limit of 110 km/h. In the controller area network data, there is a separate 'sleepiness experiment' variable, which indicates when the experiment was ongoing. The data from the turn at the midpoint was removed, as well as driving to and from the starting point. Incomplete epochs were also discarded in my analysis.
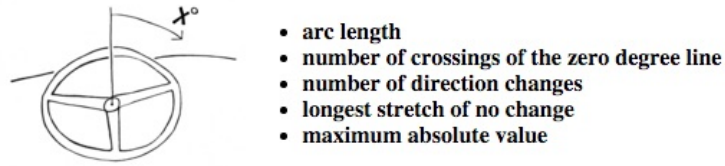
Figure 14: Summary statistics derived from the steering wheel graph

The steering wheel angle was measured in degrees with the precision of one hundredth of a degree. The first five summary statistics I created are based on this data. The first one of these five is the arc length. To estimate the length of each segment, I resorted to the general approach of using the pythagorean theorem to estimate the length of the steering wheel angle graph or two-hundred steps.

The second summary statistic based on data about the steering wheel angle is the number of crossings of the zero degree line during one twenty second epoch. This summary statistic is created by counting how many times the sign of the steering wheel angle changes during the two-hundred measurements in the epoch.

$$d = \begin{cases} 0 & \text{if } V_1 = V_0 \\ 1 & \text{if } V_1 > V_0 \\ -1 & \text{if } V_1 < V_0 \end{cases} \qquad (1)$$

The third summary statistic based on steering wheel angle data is the number of direction changes in an epoch. This summary statistic relies on the getDirection, that calculates a direction d for the change in steering wheel angle. It compares the two latest measured values V_1 and V_0 and the summary statistic is the number of changes is the d value during the twenty-second epoch. As noted earlier, previous research suggests that when people grow more tired, they make fewer and larger corrections.

A fourth summary statistic created on the basis of data about the steering wheel angle is the longest stretch with no change in value. A lack of attention while steering is a symptom of drowsiness that has also been studied in previous research. The summary statistic I created simply compares each value to the previous value, adds up the stretch, and then returns the longest stretch for each epoch.

The fifth and final summary statistic based on steering wheel angle data is the maximum absolute value of the steering wheel angle for each epoch. This summary statistic can also reveal drowsiness by showing if and when the driver made sudden large corrections. I attempted to capture all possible ways of detecting sudden steering movements or moments when the driver was lacking attention.

- speed change within an epoch
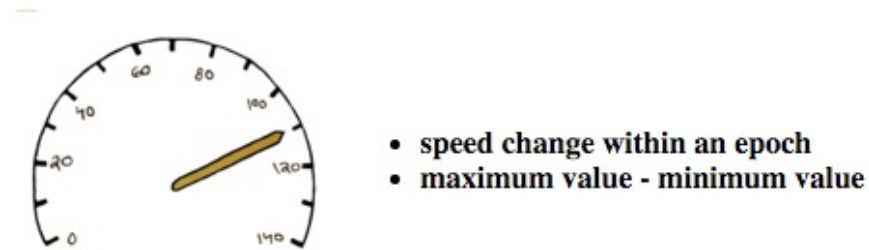- maximum value - minimum value

Figure 15: Summary statistics derived from vehicle speed graph

A second source of data used in this study for the creation of summary statistics was the speedometer. In Åkerstedt's study [3] speed was measured

in km/h, with the precision of ten meters per hour. The change in vehicle speed was recorded for each epoch, as well as the range of values. These, along with the following statistics that were created with data concerning the accelerator pedal, attempt to capture signs of drowsiness that result in altered behaviour in the use of the accelerator pedal, such as when the driver tries to compensate for their perceived drowsiness by either increasing or decreasing speed.
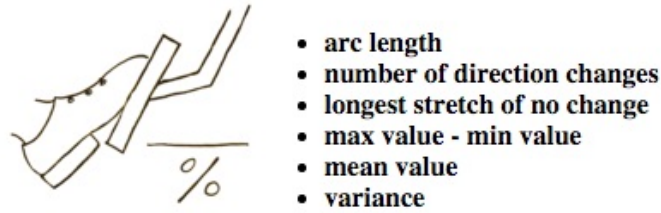
Figure 16: Summary statistics derived from the accelerator pedal graph

The rest of the summary statistics concerning the accelerator pedal are based on the fraction of how much the accelerator pedal was pressed. The precision is one tenth of a percentage point, and the value one hundred corresponds to the pedal being fully pressed. I used summary statistics for arc length, number of direction changes, longest stretch of no change, and for the range of values just like before. In addition to those there are variables for mean position, and variance.

4.3   Training the classifier

The most important element in the solution I propose for predicting which drivers can't complete their night-time drive is the classifier. Supervised machine learning works by showing labelled data from which the classifier learns the integral features on the basis of which to classify the data. It is vital to show the classifier enough samples to learn about all kinds of situations, but the aim is that we don't need to label every instance, but the machine can learn to classify the data.

I trained the classifier to detect drowsiness from individual epochs by showing the epochs alongside their manually scored drowsiness values. After this, the classifier scores new unseen epochs and those predictions are used

to solve the main problem, which was basically to identify those drivers who would not complete their drives the following night. The overview of the solution is shown in Fig. 17.

I used three commonly used classifiers. The first one of these was the nearest neighbour classifier [18]. The nearest neighbour classifier stores data points from a training set along with their corresponding labels. Here, the data used consisted of the thirteen summary statistics I created, and the label was whether any physical signs of drowsiness had been observed during the epoch concerned. For new unseen data, the nearest k neighbours were viewed, and a majority vote was used to decide the label for that data. I used $k = 1$ as number of relevant neighbours, and the closeness was measured using the Euclidean distance.

The second classifier I used is the naive Bayes' classifier, which relies on the Bayes' theorem with an expectation of strong independence of variables. This is especially naive here, where most variables are really strongly dependent. It is obvious that the position of the accelerator pedal affects the vehicle speed, and if the steering wheel remains steady for twenty seconds, the length of the curve and the number of direction changes during that time will both be zero.

The third classifier I used is the logistic regression [19]. Classifier tries to learn a regression curve as to how the different variables relate to drowsiness. It's a special case of linear regression, which is about finding a regression line that best describes the correlation between variables. A commonly used example for this is the relationship between the number of exercises that a student does during a school course and the probability that they will pass the course's final exam. Some students can pass even though they do very few exercises, and all those who do every exercise are very likely to pass.

In addition to the choice for a classifier is the choice of the training set. It can't include any samples of the set that is going to be classified, and clearly the choice of the training set has a huge impact on the performance of the classifier. The common practice is to split data set into two groups. One to train the classifier on and one set to test the performance of the classifier on. In my study, I created a training set using only epochs from the night-time drive, since I wanted to include all of the epochs from the day-time drives in validating the classifier and to use the same training set for each driver.

I tried all possible combinations of training sets up to seven members

Training the classifier

Step 1:

Each drive is split into 20 second epochs.

Step 2:

Each 20s epoch is visually scored for signs of drowsiness. Score is 0 for no signs, 1 for any signs.

Step 3:

13 summary statistics are derived from steering wheel angle, accelerator pedal and speed graphs.

Step 4:

The drowsiness score and the 13 summary statistics from each epoch are shown to the classifier.

Classification

Step 5:

Each unseen drive is split into 20 second epochs.

Step 6:

Only the 13 statistics are shown to the classifier, which gives its estimate of the drowsiness.

Step 7:

These estimates are tallied and the number of drowsy epochs is divided by the number of epochs in a drive to get a ratio of drowsy epochs.

Figure 17: Two phases of classification: training the classifier and classifying new data.

to see if any of these could perfectly predict how the drivers would do in their following night's drive. This can offer some evidence as to how well the chosen classifiers could perform. Training the classifier is the most critical part of the process since the examples shown to the classifier must contain samples of the features we want the classifier to detect.

Full daytime drive data is classified using the classifier. I then used the ratio of drowsy epochs to predict the drivers' following night-time performance. It was not obvious what the size of the training set should be, since I had no clear idea about how to classify the twenty second epochs or about what kinds of events I would be looking from the data.

For the nearest neighbour classifier, I chose four drivers that completed the whole drive and three that had to discontinue their driving either by their own choice or on the request of the accompanying test leader. Drivers one, three, seven and eleven completed their drives, while drivers ten, twelve and thirteen discontinued. This combination was chosen because it gave the best separation for the two groups.

In total, I had 1308 epochs for the best nearest neighbour classifier. Out of these, 185 contained drowsiness and 1123 didn't. From each epoch, the thirteen summary statistics were first calculated from the steering wheel angle, accelerator pedal position and vehicle speed graphs. I showed the classifier these statistics and also whether the person scoring the drive had considered this epoch to contain physical signs of drowsiness or not.

The first training set contained data about three completed drives (by drivers 1, 5 and 7) and about one drive that had to be discontinued (by driver 13). There were 801 epochs in total, out of which 60 contained signs of drowsiness and 741 were alert. This training set was used with Naive Bayes' to produce the best classification result.

The second training set contained data about three successful night-time drives (by drivers 1, 7 and 15) and about one discontinued night-time drive (by driver 20). This training set performed best for the logistic regression. In this training set, there were 706 epochs containing only alert driving and 148 epochs containing physical signs of drowsiness.

KDS was used as a marker for drowsiness, and all non-zero values were truncated into a single value of 1 for epochs that show any amount of alertness. KDS was chosen because the drivers who did not finish their drives had significantly higher KDS values than drivers who completed both of their drives. KDS has previously been validated as a reliable means of measuring drivers' drowsiness. This was confirmed by a study where high KDS was proven to connect with increased lane departures as well as higher mean and variability of blink duration [5].

4.4   Results

All drivers completed their daytime drives. Five of the eighteen participating drivers had their drives terminated by the test leader accompanying them in the car. Two drivers decided to stop driving because they felt that their safety was compromised, and one driver requested a break which resulted

in the test leader terminating the drive. In total, ten of eighteen completed their night-time drive and eight didn't. Drivers four and six cancelled their participation late and could not be replaced in time.

The following Tables are all in the same format. Drivers are grouped into completing (C) and not-completing (NC) drivers based on whether they could finish the night-time drive or not. Column 'Alert' has the number of epochs that were labelled as alert by the classifier. 'Drowsy' contains the number of epochs labelled as drowsy by the classifier in question. 'Ratio' contains the number of drowsy epochs divided by the total number of epochs in the drive.

The figures contain the ratios given by each classifier sorted from lowest to highest. Figure 18 shows the classification when the human-scored values were reduced to 0 or 1. The simplification was done by giving the epoch a score of 0 if there was no signs of drowsiness and 1 if there were any signs.

Figure 19 shows how nearest neighbour classifier is able to separate four of the eight drivers who fail to complete their nighttime drives. These drivers are distinguished from the rest of the drivers with a clear margin. Each daytime drive from the study is used to train the classifier that is used to classify all of the drives.

Results shown in Figs. 20 to 22 are suspect to overfitting because of the choice of the training set. All daytime drives in each case were classified using the same classifier. The training set used in training, however, is the best possible one for each classifier. This was done to show that a carefully trained classifier could be used to detect drowsiness even from previously unseen drivers.

Naive Bayes' and logistic regression classifiers were also trained with full night-time data. The imbalance of alert and drowsy samples in the training data causes them to classify nearly every epoch as alert. The usual remedy for this is to randomly sample the training data to select a balanced training set. I chose not to, because I was reluctant to introduce randomness into my research.
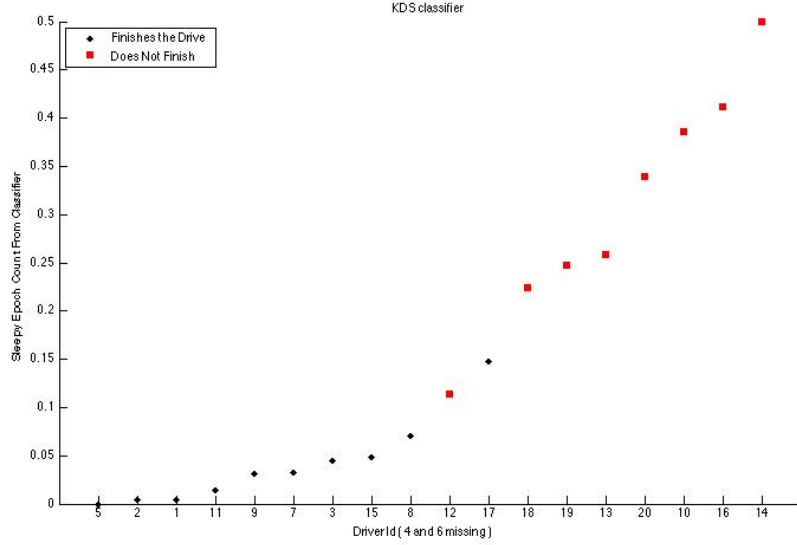
Figure 18: Classification using the manually scored KDS

Table 4 and Fig. 18 show the results, when actual manually scored KDS values were used to detect drowsiness. Only driver 17, who completed both drives, is given a higher ratio of epochs with drowsiness compared to driver 12, who is scored with the least amount of drowsiness among the drivers who couldn't complete their night-time drive. This is in line with the suggestion of the original research group [3], that it would be possible already after the daytime drive to identify those drivers who would not complete their following night-time drive.

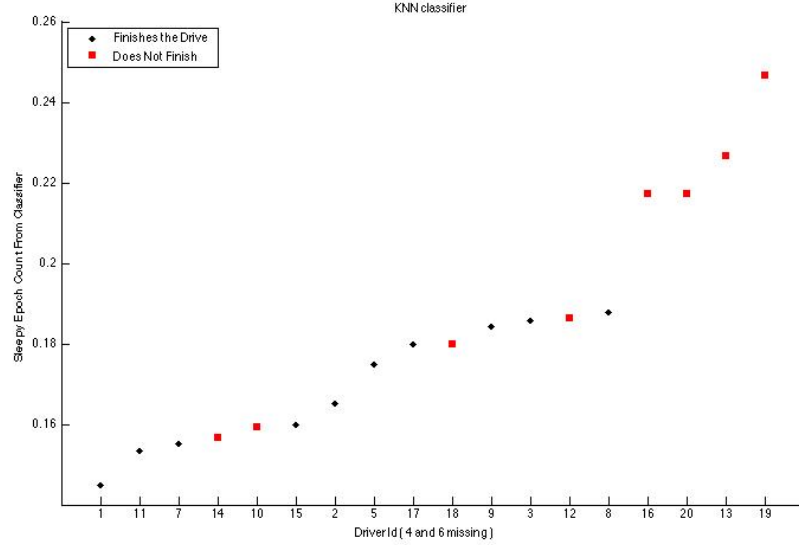| Id	Group	Alert	Drowsy	Ratio
5	C	223	0	0.000
1	C	227	1	0.004
2	C	229	1	0.004
11	C	212	3	0.014
9	C	221	7	0.031
7	C	212	7	0.032
3	C	216	10	0.044
15	C	214	11	0.049
8	C	213	16	0.070
12	NC	195	25	0.114
17	C	185	32	0.147
18	NC	177	51	0.224
19	NC	171	56	0.247
13	NC	167	58	0.258
20	NC	146	75	0.339
10	NC	139	87	0.385
16	NC	130	91	0.412
14	NC	115	115	0.500

Table 4: KDS is used to classify the daytime drives

43

Figure 19: K nearest neighbour classification with k=1 and all night-time drives as a training set

Table 5 and Fig. 19 show the results for k nearest neighbour classifier with k=1 and with all of the night-time drives as a training set. Drivers sixteen, twenty, thirteen and nineteen stand out from the rest of the drivers. Those four drivers are also among the six drivers scored most drowsy by KDS.

Drivers ten and fourteen, who were among the three with most signs of drowsiness, were labelled with really low drowsiness. Driver five, who showed no signs of alertness during the day, was scored with thirty-nine epochs of drowsiness, while drivers one and two, each with only one epoch of drowsiness during the day, got thirty-tree and thirty-eight drowsy epochs from the classifier, respectively.

Id	Group	Alert	Drowsy	Ratio
1	C	228	33	0.145
11	C	215	33	0.153
7	C	219	34	0.155
14	NC	230	36	0.157
10	NC	226	36	0.159
15	C	225	36	0.160
2	C	230	38	0.165
5	C	223	39	0.175
17	C	217	39	0.180
18	NC	228	41	0.180
9	C	228	42	0.184
12	NC	220	41	0.186
3	C	226	42	0.186
8	C	229	43	0.188
16	NC	221	48	0.217
20	NC	221	48	0.217
13	NC	225	51	0.227
19	NC	227	56	0.247

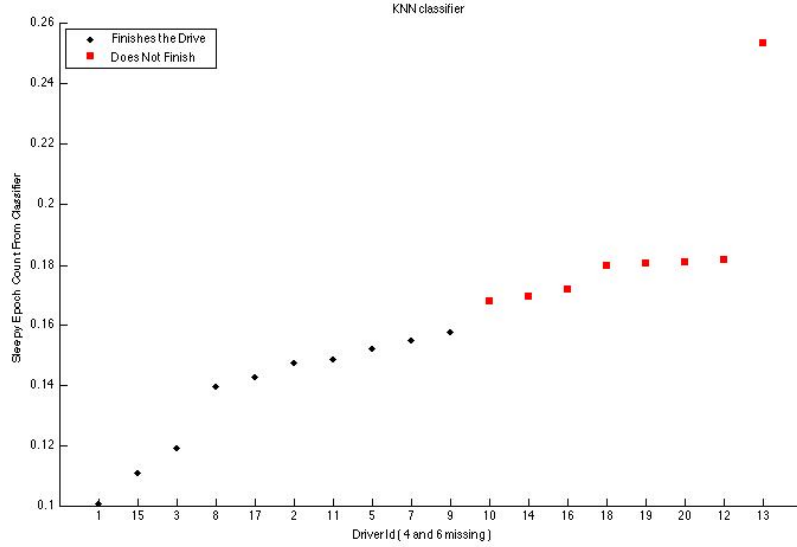Table 5: K nearest neighbour classification with k=1 and all night drives as a training set

Figure 20: K nearest neighbour classification with k=1 and drivers 1, 3, 7, 10, 11, 12, and 13 as a training set

Table 6 and Fig. 20 show the results of the nearest neighbour classifier, trained the with nighttime drives of drivers one, three, seven, ten, eleven, twelve and thirteen, with $k = 1$ to classify the daytime drives. The results were achieved using the brute force method of checking all possible combinations of training sets up to group size of seven.

This classifier split the completing and terminating drivers perfectly with the highest ratio of drowsiness by the classifier for the completing drivers being 0.158 and and the lowest one for the terminating drivers being 0.168. The method of brute forcing the training sets was chosen first of all to see if any of the combinations produced useful classifiers, and secondly, to study how the performance of the classifier could be measured.

Id	Group	Alert	Drowsy	Ratio
1	C	205	23	0.101
15	C	200	25	0.111
3	C	199	27	0.119
8	C	197	32	0.140
17	C	186	31	0.143
2	C	196	34	0.148
11	C	183	32	0.149
5	C	189	34	0.152
7	C	185	34	0.155
9	C	192	36	0.158
10	NC	188	38	0.168
14	NC	191	39	0.170
16	NC	183	38	0.172
18	NC	187	41	0.180
19	NC	186	41	0.181
20	NC	181	40	0.181
12	NC	180	40	0.182
13	NC	168	57	0.253

Table 6: K nearest neighbour classification with k=1 and drivers 1, 3, 7, 10, 11, 12, and 13 as a training set
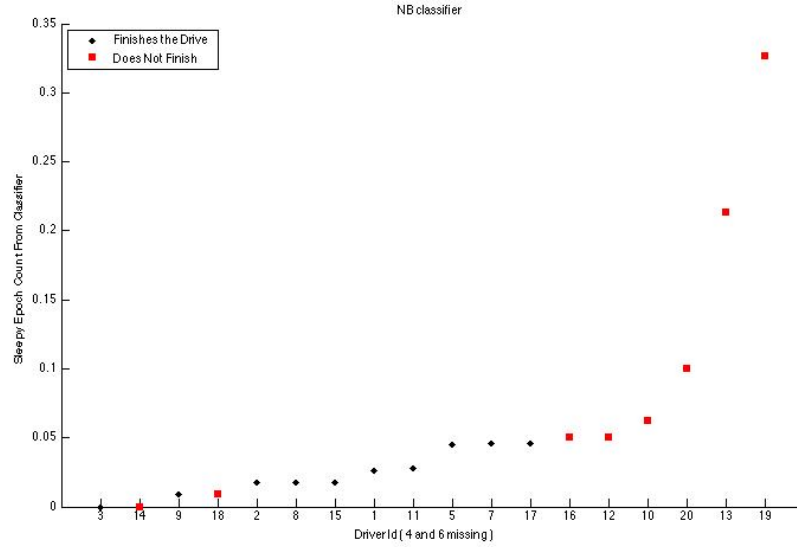
45

Figure 21: Naive Bayes' classification with drivers 1, 5, 7, and 13 as a training set

Table 7 and Fig. 21 show the results when naive Bayes' classifier was trained with a training set based on data from the night-time drives of drivers one, five, seven and thirteen. This classifier rates all but two not-completing drivers with higher drowsiness than the completing drivers. Like the previous result, this was achieved using the brute force method.

Drivers twenty, thirteen and nineteen stand out from these results quite nicely. They were also among the highest rated when KDS was used to classify the drivers. Drivers ten and sixteen, who showed a lot of drowsiness, were scored with fourth and fifth most drowsiness by the classifier, but driver number fourteen, who showed the most drowsiness when scored by a human, was scored with perfect alertness by the Naive Bayes' classifier.

Id	Group	Alert	Drowsy	Ratio
3	C	226	0	0.000
14	NC	230	0	0.000
9	C	226	2	0.009
18	NC	226	2	0.009
2	C	226	4	0.017
8	C	225	4	0.017
15	C	221	4	0.018
1	C	222	6	0.026
11	C	209	6	0.028
5	C	213	10	0.045
7	C	209	10	0.046
17	C	207	10	0.046
12	NC	209	11	0.050
16	NC	210	11	0.050
10	NC	212	14	0.062
20	NC	199	22	0.100
13	NC	177	48	0.213
19	NC	153	74	0.326

Table 7: Naive Bayes' classification with drivers 1, 5, 7, and 13 as a training set
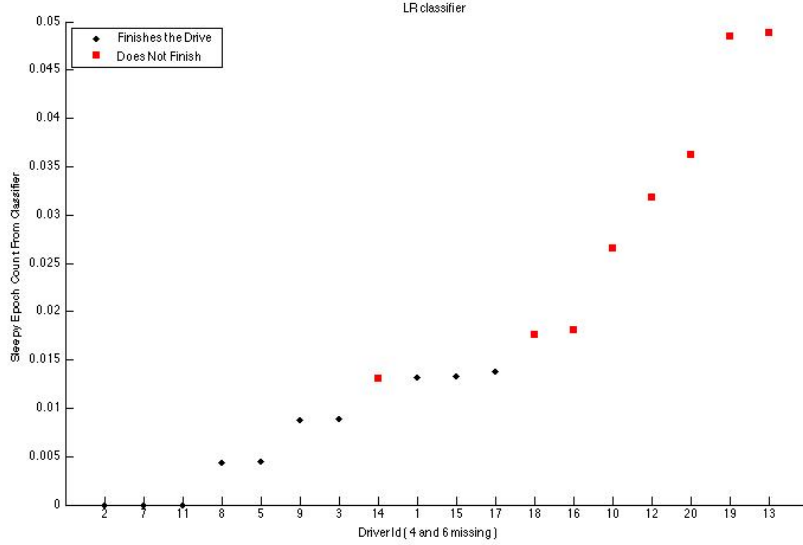
Figure 22: Logistic regression classification with drivers 1, 7, 15, and 20 as a training set

Table 8 shows the results when the logistic regression classifier was trained with a training set based on the data from the night-time drives of drivers one, seven, fifteen and twenty. This training set contained altogether 799 alert epochs and 94 drowsy epochs. The logistic regression classifier scored all but one not-completing driver with higher drowsiness than any of the completing drivers.

The performance was evaluated by judging how good of a split the classifier produces and not by how well individual epochs are labelled. Even the driver with the highest rated drowsiness is scored with eleven epochs

Id	Group	Alert	Drowsy	Ratio
2	C	230	0	0.000
7	C	219	0	0.000
11	C	215	0	0.000
8	C	228	1	0.004
5	C	222	1	0.004
9	C	226	2	0.009
3	C	224	2	0.009
14	NC	227	3	0.013
1	C	225	3	0.013
15	C	222	3	0.013
17	C	214	3	0.014
18	NC	224	4	0.018
16	NC	217	4	0.018
10	NC	220	6	0.027
12	NC	213	7	0.032
20	NC	213	8	0.036
19	NC	216	11	0.048
13	NC	214	11	0.049

Table 8: Results from logistic regression classifier

of drowsiness, while driver fourteen has half of their epochs labelled as drowsy by a human researcher observing their EEG and eye movement.

The main result in my work is the one from the nearest neighbour classifier

trained with full night-time data. The results from the classification are shown in Table 5 and in Fig. 19. That classifier was trained with a full data set and it achieved a result that compared very favourably to the results presented in the Section 3.3. The four drivers were scored with a ratio of 0.217, 0.217, 0.227 and 0.247 when the next highest ratio was 0.188.

Other results were achieved by hand-picking the training set. This was done in order to show that the problem could be solved with a proper choice of the training set. That the two groups of drivers who did and didn't complete their night-time drives could be separated by a classifier that has been trained to mimic the performance of the human expert, looking for the signs of drowsiness from the EEG and from the electrooculograph.

When the naive Bayes' and logistic regression classifiers were trained with the full data set they were overwhelmed with the number of alert samples in the training data. Naive Bayes' classifier trained with a full data set labelled only four drivers with more than two epochs with drowsiness, or a ratio of more than 0.01. One driver was classified with 34 epochs of drowsiness and received a ratio of 0.15 drowsy epochs.

The logistic regression classifier, trained with a full data set, labelled only three epochs from driver nineteen as drowsy and all other epochs were classified as alert. Curiously driver nineteen was also the same driver, who was labelled with 34 drowsy epochs by the naive Bayes' classifier.

5   Conclusions

Driving tired remains a huge problem. The best summary of the issue is the fact that one driver, who admitted to falling asleep did complete the drive and neither the driver nor the test leader, whose only job was to monitor the alertness of the driver did terminate the drive.

Chapter two showed what is the basis for the tests used as a baseline when developing new methods to detect drowsiness. There are many applications for machine learning in that field as well. We live in the age of biohacking and recording all kinds of data from oneself. It's not that far-fetched idea that people would start wearing some kind of EEG device or some an eyewear, which would record movement of the eyes at all times.

Chapter three presented the current state of detecting drowsiness of the driver. It seems that there are no clear winner as to which signal should be

used and how it should be analysed. The current solutions show promise but like Fig. 11 show, most solutions do not differ from others when it comes to their predicting power.

5.1 Discussion

A lot of research has been conducted on studying drowsiness. Sleep-related problems can cause trouble in a variety of other fields. There are jobs that require staying alert for extended periods of time, like flying a plane or political negotiations. Many medical problems can cause drowsiness, and being able to measure and detect the effects different medicines and their combinations could help patients a lot. Self-reporting is often used as a method in studies focusing on drowsiness, but at the same time, 26% of drivers reported having fallen asleep at the wheel [49].

While people can be adept at sensing their own drowsiness, they don't necessarily experience a sharp decline in their capabilities, but rather drift slowly off to sleep. Knowing this can help with future research of drowsiness metrics and must be considered as a huge need for such metrics because while people know they are tired they do not sense imminent sleep.

The effects of the circadian rhythm and of staying awake for extended periods of time are already somewhat understood, but a lot of interesting new research on these is also currently underway. For example, in sports-related studies, measurements are made on how performance changes during the day and how it is affected by sleep. This kind of research could help separate normal events from drowsiness events if the circadian rhythm and the time of the day are used when looking at driving data.

Elon Musk recently described in an interview the 'fleet learning mechanism' [70] with the aim of developing a self-driving car. According to him, the 100000 Tesla Model S cars, that are currently in use, are driven for about 1.5 million miles each day with various devices recording their activity. The recorded data is sent to servers [70], providing so much data for the company that it can be used much like the black boxes in airplanes to piece together events that lead to accidents.

The biggest lack in trying to develop reliable drowsiness detectors on the basis of this type and amount of data is the lack of reliable performance measures. If it was possible to define what sort of behaviour should be expected of drivers, then their actual driving behaviour could be compared

against that definition. The nearly billion kilometres of driving recorded with Tesla Model S cars allows for a completely different kind of research.

Finding convincing biomarkers for drowsiness is an active field of study. A variety of different approaches have already been tested. What I find comforting is the **flood of data that will become available** as cars are made to be more sensitive of their surroundings. Developing self-driving cars and new features like parking assistant will require for cars to also have more "self-awareness". This future influx of data can help researchers if only the proper tools can be found for dealing with such amounts of data.

Eighteen ninety-minute drives do clearly not produce enough data for creating a universal classifier for studying drowsy driving. However, the sample sizes of many other studies on the same subject are very similar, suggesting that studying drowsiness is rather expensive and time consuming.

Detecting drowsiness is really hard. Even in laboratory settings there is a lot of variation between tested subjects, and experts do not always agree on how to interpret the results. My study suggests, however, that the human scored KDS score of drowsiness could have potential as a means to predict driving performance when it comes to tiredness. It is not out of the question that that a machine could be trained to do this as well.

The nearest neighbour classifier with full data as a training set was able to identify with a clear margin half of those drivers who failed to complete their following night-time drive. This gives a lot of hope for believing that ways to prevent drowsy driving could be developed with the aid of machine learning. The fact that the same classifier was used for all drivers is also promising. Even if the machine used would require calibration, it could serve in law enforcement if the calibration was done as part of the process required to receive a driving permit.

The biggest problem in the present study is probably the need to resort to hand-picked training sets when looking for better performance than the fourteen out of eighteen score of the nearest neighbour classifier used with full night-time drive data set. Another issue is the poor performance of individual KDS prediction. The research problem was not to replicate the KDS prediction so not much effort was given to that problem, but the fact that the driver who showed the most drowsiness in EEG was classified as completely alert by trained classifier is an issue that requires further attention.

5.2   Summary of Contributions

This chapter lists the contributions of this work and the insight gained from this research. The contributions are numbered from most to least important.

1. I showed that the nearest neighbour classifier can detect four of the eight non-completing drivers with confidence and without any false positive cases. The lack of false positives is especially significant since people tend to disregard any signal that alerts when there is no need for it. Also the clear margin by which the four clear cases stand out from the rest gives some backing to these results.

2. A hand-picked training set could perfectly predict which drivers would be able to complete their night-time drives simply by looking at the three selected sources of data. These three data sources were the steering wheel angle, vehicle speed and the accelerator pedal position. These results come from a very narrow metrics of success and performance for the classifier and this would be an excellent point of focus for future research.

3. Each machine learning tool that was tried was able to detect at a reasonably good rate which of the drivers would not be able to complete their night-time drives. While the rate of reliability could also have been better, this nevertheless gives reason to believe that the results were not a fluke, but that drowsiness can indeed be detected at least to some extent on the basis of measurements made on the car and not on the driver.

Other point worth noting is that counting epochs that have at least some amount of drowsiness in them could be a better predictor for drowsiness than trying to detect the level of drowsiness in each twenty second epoch. This claim is supported by what Åkerstedt's research group stated themselves about not finding any interesting signs of different behaviour just prior to the termination of the drive. Studies suggest that a human's level of alertness fades slowly instead of crashing suddenly. This should be taken into consideration in future research.

Machine learning can become a useful tool for detecting drivers' drowsiness. When we are able to access the huge amounts of data that car companies and insurance companies are currently collecting, computer-based learning can be used to learn to spot the pattern leading up to an event caused by drowsiness. The amount of data is such that it would be impossible to use humans to comb through it and search the signs of drowsiness.

5.3 Future Research

Machine learning is an useful tool for each aspect of this work. It can help with detecting sleep spindles and other changes in the brain waves and even in speech. Many aspects of driving are studied in both laboratory and real world settings and Internet of Things will allow for smaller sensors and better data collection. Already many cars are highly sensitive to their surroundings.

Next to no effort was spent on feature selection and to a critical review of the sources of data. Are these three the best predictors of data? What about the thirteen summary statistics? This data set is really good for studying individual KDS prediction performance. Daytime drives have 3392 alert epochs and 646 drowsy epochs in total . Night-time drives have 2877 and 634 alert and sleepy epochs respectively. Finding a way to replicate human expert performance as drowsiness detector would be a great service to many fields of study.

The second path for future research is the development of goodness measure for cars. The Washington Post [29] reported that the insurance companies are willing to lower their fees if the drivers download an application that measures the breaking and acceleration of the car. This data is then used to predict the probability of an accident. Clearly, the same info is relevant here. When a driver is not alert, their reactions become more sudden.

Another source of data could be taxi drivers and bus drivers. These cars already have computers for the driver, and I imagine that a reasonably sensitive sensor pack could be made and attached to each car to measure driving data. The drowsiness label could be gained from the drivers by using self-reported drowsiness or by checking how many hours that person has been working already.

The data set also includes the EEG signal and the drowsiness score given by a human expert for each twenty second epoch of each drive. One possible direction of study could be the automatic labelling of the EEG data, or replicating the performance of the expert. This is a more traditional classification problem, than trying to detect non-completing drivers from a group of eighteen drivers.

The best part in this is that even small gains in detecting drowsiness can help save human lives.

References

[1] ÅKERSTEDT, T., FOLKARD, S., AND PORTIN, C. Predictions from the three-process model of alertness. *Aviation, Space, and Environmental Medicine 75*, Supplement 1 (2004), A75–A83.

[2] ÅKERSTEDT, T., AND GILLBERG, M. Subjective and objective sleepiness in the active individual. *International Journal of Neuroscience 52*, 1-2 (1990), 29–37.

[3] ÅKERSTEDT, T., HALLVIG, D., ANUND, A., FORS, C., SCHWARZ, J., AND KECKLUND, G. Having to stop driving at night because of dangerous sleepiness–awareness, physiology and behaviour. *Journal of Sleep Research 22*, 4 (Aug 2013), 380–388.

[4] AMERICAN ACADEMY OF SLEEP MEDICINE, CHICAGO, ILLINOIS. International classification of sleep disorders, revised: Diagnostic and coding manual, 2001.

[5] ANUND, A., KECKLUND, G., PETERS, B., FORSMAN, Å., LOWDEN, A., AND ÅKERSTEDT, T. Driver impairment at night and its relation to physiological sleepiness. *Scandinavian Journal of Work, Environment & Health 34*, 2 (2008), 142–150.

[6] ARONSON, J. K. Biomarkers and surrogate endpoints. *British Journal of Clinical Pharmacology 59*, 5 (May 2005), 491–494.

[7] BASNER, M., MOLLICONE, D., AND DINGES, D. F. Validity and sensitivity of a brief psychomotor vigilance test (pvt-b) to total and partial sleep deprivation. *Acta Astronaut 69*, 11-12 (Dec 2011), 949–959.

[8] BAYES, T., AND PRICE, R. An essay towards solving a problem in the doctrine of chances. by the late rev. mr. Bayes, F.R.S. communicated by mr. Price, in a letter to John Canton, A.M.F.R.S. *Philosophical Transactions (1683-1775) 1763* (370–418).

[9] BENBADIS, S. R. Introduction to sleep electroencephalography. *Sleep: A Comprehensive Handbook. USA: John Wiley & Sons* (2006), 989–1024.

[10] BERGER, H. Über das elektrenkephalogramm des menschen. *European Archives of Psychiatry and Clinical Neuroscience 87*, 1 (1929), 527–570.

[11] BREIMAN, L. Bagging predictors. *Machine learning 24*, 2 (1996), 123–140.

[12] BREIMAN, L. Arcing the edge. Tech. rep., Technical Report 486, Statistics Department, University of California at Berkeley, 1997.

[13] BREIMAN, L. Random forests. *Machine learning 45*, 1 (2001), 5–32.

[14] BROWN, M., MARMOR, M., VAEGAN, ZRENNER, E., BRIGELL, M., BACH, M., AND ISCEV. Iscev standard for clinical electro-oculography (eog) 2006. *Documenta Ophthalmologica 113*, 3 (Nov 2006), 205–12.

[15] CAFFIER, P. P., ERDMANN, U., AND ULLSPERGER, P. Experimental evaluation of eye-blink parameters as a drowsiness measure. *European journal of applied physiology 89*, 3-4 (May 2003), 319–25.

[16] CARSKADON, M. A., DEMENT, W. C., MITLER, M. M., ROTH, T., WESTBROOK, P. R., AND KEENAN, S. Guidelines for the multiple sleep latency test (mslt): A standard measure of sleepiness. *Sleep 9*, 4 (1986), 519–524.

[17] CHERRY, E. C. Some experiments on the recognition of speech, with one and with two ears. *The Journal of the acoustical society of America 25*, 5 (1953), 975–979.

[18] COVER, T. M., AND HART, P. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory 13*, 1 (January 1967), 21–27.

[19] COX, D. R. The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological)* (1958), 215–242.

[20] DE MAIRAN, J. Observation botanique. *Histoire de l'Académie royale des sciences avec les mémoires de mathématique et de physique tirés des registres de cette Académie 35* (1729), 36.

[21] DIJK, D.-J., DUFFY, J. F., AND CZEISLER, C. A. Circadian and sleep/wake dependent aspects of subjective alertness and cognitive performance. *Journal of Sleep Research 1* (1992), 112–117.

[22] DINGES, D. F., AND POWELL, J. W. Microcomputer analyses of performance on a portable, simple visual rt task during sustained operations.

Behavior Research Methods, Instruments, & Computers 17, 6 (1985), 652–655.

[23] EDWARDS, B., WATERHOUSE, J., AND REILLY, T. The effects of circadian rhythmicity and time-awake on a simple motor task. *Chronobiology International 24*, 6 (2007), 1109–1124.

[24] EDWARDS, B., WATERHOUSE, J., AND REILLY, T. Circadian rhythms and their association with body temperature and time awake when performing a simple task with the dominant and non-dominant hand. *Chronobiology International 25*, 1 (2008), 115–132.

[25] ENDSLEY, M. R. Towards a theory of situation awareness in dynamic systems. *Human Factors 37*, 1 (1995), 32–64.

[26] FAHY, F. J. Measurement of acoustic intensity using the cross-spectral density of two microphone signals. *The Journal of the Acoustical Society of America 62*, 4 (1977), 1057–1059.

[27] FOR DISEASE CONTROL, C., AND PREVENTION. Unhealthy sleep-related behaviors — 12 states, 2009. *Morbidity and Mortality Weekly Report 60*, 8 (March 2011), 1–6.

[28] FORSMAN, P. M., VILAA, B. J., SHORTC, R. A., MOTT, C. G., AND VAN DONGEN, H. P. Efficient driver drowsiness detection at moderate levels of drowsiness. *Accident Analysis and Prevention 50* (2013), 341–350.

[29] FUNG, B. The big data of bad driving, and how insurers plan to track your every turn. *The Washington Post*, 4 (2016).

[30] GILLBERG, M., KECKLUND, G., AND ÅKERSTEDT, T. Sleepiness and performance of professional drivers in a truck simulator—comparisons between day and night driving. *Journal of Sleep Research 5*, 1 (1996), 12–15.

[31] GOEL, N., BASNER, M., RAO, H., AND DINGES, D. F. Circadian rhythms, sleep deprivation, and human performance. *Progress in molecular biology and translational science 119* (2013), 155–190.

[32] GRAW, P., KRÄUCHI, K., KNOBLAUCH, V., WIRZ-JUSTICE, A., AND CAJOCHEN, C. Circadian and wake-dependent modulation of fastest and slowest reaction times during the psychomotor vigilance task. *Physiology & Behavior 80* (2004), 695–701.

[33] HIGUCHI, T. Approach to an irregular time series on the basis of the fractal theory. *Physica D: Nonlinear Phenomena 31*, 2 (1988), 277–283.

[34] HODDES, E., ZARCONE, V., SMYTHE, H., PHILLIPS, R., AND DEMENT, W. Quantification of sleepiness: a new approach. *Psychophysiology 10*, 4 (1973), 431–436.

[35] HUNT, E. B., MARIN, J., AND STONE, P. J. Experiments in induction. *Academic Press* (1966).

[36] JAMMES, B., SHARABTY, H., AND ESTEVE, D. Automatic eog analysis: A first step toward automatic drowsiness scoring during wakesleep transitions. *Somnologie 12* (2008), 227–232.

[37] JOHNS, M. W. A new method for measuring daytime sleepiness: the epworth sleepiness scale. *Sleep 14*, 6 (1991), 540–545.

[38] JUTTEN, C., AND HERAULT, J. Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture. *Signal processing 24*, 1 (1991), 1–10.

[39] KOHONEN, T. The self-organizing map. *Neurocomputing 21*, 1 (1998), 1–6.

[40] KOISAARI, T. Valt-vuosiraportti 2013. Tech. rep., Vakuutusyhtiöiden liikenneturvallisuustoimikunta VALT, 2014.

[41] KRAJEWSKI, J., BATLINER, A., AND GOLZ, M. Acoustic sleepiness detection: framework and validation of a speech-adapted pattern recognition approach. *Behavior research methods 41*, 3 (Aug 2009), 795–804.

[42] LENNÉ, M. G., TRIGGS, T. J., AND REDMAN, J. R. Time of day variations in driving performance. *Accident Analysis & Prevention 29*, 4 (1997), 431–437.

[43] LIN, C.-T., CHANG, C.-J., LIN, B.-S., HUNG, S.-H., CHAO, C.-F., AND WANG, I.-J. A real-time wireless brain–computer interface system for drowsiness detection. *IEEE Transactions on Biomedical Circuits and Systems 4*, 4 (August 2010).

[44] LIN, C.-T., WU, R.-C., LIANG, S.-F., CHAO, W.-H., CHEN, Y.-J., AND JUNG, T.-P. Eeg-based drowsiness estimation for safety driving using independent component analysis. *Circuits and Systems I: Regular Papers, IEEE Transactions on 52*, 12 (2005), 2726–2738.

[45] LIU, C. C., HOSKING, S. G., AND LENNÉ, M. G. Predicting driver drowsiness using vehicle measures: recent insights and future challenges. *Journal of safety research 40*, 4 (2009), 239–45.

[46] LOH, S., LAMOND, N., DORRIAN, J., ROACH, G., AND DAWSON, D. The validity of psychomotor vigilance tasks of less than 10-minute duration. *Behavior Research Methods 36*, 2 (2004), 339–346.

[47] LUNDIN, M., AND KANSTRUP, L. Method for detection of sleepiness: measurement of interaction between driver and vehicle. Master's thesis, Linköping University, 2006.

[48] MALTHUS, T. R. *An Essay on the Principle of Population, as it affects the Future Improvement of Society: with remarks on the Speculations of Mr. Godwin, M. Condorset, and other writers.* J. Johnson, 1798.

[49] MANSIKKA, T., AND PÖYSTI, L. Tunnetaanko liikennesäännöt? Liikenneturvan selvityksiä 2, Liikenneturva, 2013.

[50] MARDI, Z., ASHTIANI, S. N. M., AND MIKAILI, M. Eeg-based drowsiness detection for safe driving using chaotic features and statistical tests. *Journal of Medical Signals and Sensors 1*, 2 (2011), 130–137.

[51] MCCARLEY, R. W., AND SINTON, C. M. Neurobiology of sleep and wakefulness. *Scholarpedia 3*, 4 (2008), 3313.

[52] MCCULLOCH, W. S., AND PITTS, W. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics 5*, 4 (1943), 115–133.

[53] McDonald, A. D., Lee, J. D., Schwarz, C., and Brown, T. L. Steering in a random forest: Ensemble learning for detecting drowsiness-related lane departures. *Human Factors 56*, 5 (2014), 986–998.

[54] Monk, T. H. A visual analogue scale technique to measure global vigor and affect. *Psychiatry research 27*, 1 (1989), 89–99.

[55] Morris, D. M., Pilcher, J. J., and Switzer III, F. S. Lane heading difference: An innovative model for drowsy driving detection using retrospective analysis around curves. *Accident Analysis and Prevention 80* (2015), 117–124.

[56] Niedermeyer, E., and da Silva, F. L. *Electroencephalography: Basic Principles, Clinical Applications, and Related Fields.* Lippincott Williams & Wilkins; Fifth edition, 2004.

[57] Nuwer, M. R., Comi, G., Emerson, R., Fuglsang-Frederiksen, A., Guérit, J. M., Hinrichs, H., Ikeda, A., Luccas, F. J., and Rappelsburger, P. Ifcn standards for digital recording of clinical eeg. international federation of clinical neurophysiology. *Electroencephalography and clinical neurophysiology 106*, 3 (Mar 1998), 259–61.

[58] of Sleep Medicine, A. A., Iber, C., et al. *The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications.* American Academy of Sleep Medicine, 2007.

[59] Petrosian, A. Kolmogorov complexity of finite sequences and recognition of different preictal eeg patterns. In *Computer-Based Medical Systems, 1995., Proceedings of the Eighth IEEE Symposium on* (1995), IEEE, pp. 212–217.

[60] Radun, I., Radun, J., and Ohisalo, J. Väsymys tieliikenteessä ja laki. keitä rangaistaan ja milloin? Tech. rep., Liikenneturva, 2009.

[61] Rechtschaffen, A., and Kales, A., Eds. *A Manual of Standardized Terminology Techniques and Scoring System for Sleep Stages of Human Subjects.* BIS / BRI UCLA, Los Angeles, 1968.

[62] Reinberg, A., Bicakova-Rocher, A., Nouguier, J., Gorceix, A., Mechkouri, M., Touitou, Y., and Ashkenazi, I. Circadian

rhythm period in reaction time to light signals difference between right-
and left-hand side. *Cognitive Brain Research 6*, 2 (Oct 1997), 135–40.

[63] SAFETYNET. Fatigue, 2009.

[64] SEUGNET, L., BOERO, J., GOTTSCHALK, L., DUNTLEY, S. P., AND
SHAW, P. J. Identification of a biomarker for sleep drive in flies and
humans. *Proceedings of the National Academy of Sciences 103*, 52 (2006),
19913–19918.

[65] SOUSANIS, J. World vehicle population tops 1 billion units, August
2011.

[66] STERN, J. A., WALRATH, L. C., AND GOLDSTEIN, R. The endogenous
eyeblink. *Psychophysiology 21*, 1 (1984), 22–33.

[67] THIFFAULT, P., AND BERGERON, J. Monotony of road environment
and driver fatigue: a simulator study. *Accident Analysis & Prevention
35*, 3 (May 2003), 381–91.

[68] TIETÄVÄINEN, A., GATES, F. K., MERILÄINEN, A., MANDEL, J. E.,
AND HÆGGSTRÖM, E. Nintendo® wii fit based sleepiness tester detects
impairment of postural steadiness due to 24 h of wakefulness. *Medical
Engineering & Physics 35* (2013), 1850–1853.

[69] VAPNIK, V. Pattern recognition using generalized portrait method.
Automation and remote control 24 (1963), 774–780.

[70] VOELCKER, J. Tesla autopilot: The 10 most important things you
need to know. http://www.greencarreports.com/news/1100482_tesla-
autopilot-the-10-most-important-things-you-need-to-know, Oct 2015.

[71] VUCKOVIC, A., RADIVOJEVIC, V., CHEN, A. C. N., AND POPOVIC,
D. Automatic recognition of alertness and drowsiness from eeg by an
artificial neural network. *Medical engineering & physics 24*, 5 (Jun
2002), 349–60.

[72] WIDROW, B., AND HOFF, M. E. Adaptive switching circuits. *Defense
Technical Information Center* (1960).

[73] Wierwille, W., Ellsworth, L., Wreggit, S., Fairbanks, R., and Kim, C. Research on vehicle-based driver status/performance monitoring; development, validation, and refinement of algorithms for detection of driver drowsiness. Final Report DOT HS 808 247, U.S. Department of Transportation, 1994.

[74] World Health Organization. Global status report on road safety 2015. Tech. rep., World Health Organization, 2015.

[75] Wright, K. P., Hull, J. T., and Czeisler, C. A. Relationship between alertness, performance, and body temperature in humans. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology 283*, 6 (2002), R1370–R1377.

[76] Xiong, G., Zhang, L., Liu, H., Zou, H., and Guo, W. A comparative study on apen, sampen and their fuzzy counterparts in a multiscale framework for feature extraction. *Journal of Zhejiang University SCIENCE A 11*, 4 (2010), 270–279.

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