Proceedings of the Human Factors and Ergonomics Society Annual Meeting

http://pro.sagepub.com/

Real-Time Detection of Drowsiness Related Lane Departures Using Steering Wheel Angle

Anthony D. McDonald, Chris Schwarz, John D. Lee and Timothy L. Brown

Proceedings of the Human Factors and Ergonomics Society Annual Meeting 2012 56: 2201
DOI: 10.1177/1071181312561464

Real-Time Detection of Drowsiness Related Lane Departures Using Steering Wheel Angle

Anthony D. McDonald[1], Chris Schwarz[2], John D. Lee[1], Timothy L. Brown[2]
[1]University of Wisconsin-Madison, Madison, WI, USA
[2]National Advanced Driving Simulator, Iowa City, IA, USA

Drowsy driving is a significant factor in many motor vehicle crashes in the United States and across the world. Efforts to reduce these crashes have developed numerous algorithms to detect both acute and chronic drowsiness. These algorithms employ behavioral and physiological data, and have used different machine learning techniques. This work proposes a new approach for detecting drowsiness related lane departures, which uses unfiltered steering wheel angle data and a random forest algorithm. Using a data set from the National Advanced Driving Simulator the algorithm was compared with a commonly used algorithm, PERCLOS and a simpler algorithm constructed from distribution parameters. The random forest algorithm had higher accuracy and Area Under the receiver operating characteristic Curve (AUC) than PERCLOS and had comparable positive predictive value. The results show that steering-angle can be used to predict drowsiness related lane-departures six seconds before they occur, and suggest that the random forest algorithm, when paired with an alert system, could significantly reduce vehicle crashes.

 DOI 10.1177/1071181312561464

INTRODUCTION

The National Highway Traffic Safety Administration (NHTSA) reports that drowsiness "causes more than 100,000 [automobile] crashes a year, resulting in 40,000 injuries, and 1,550 deaths" (National Highway Traffic Safety Administration, 2011). Because drowsiness leaves no physical trace, these statistics might underestimate the true magnitude of the problem. The 100-Car naturalistic driving study found that drowsy driving contributed to 22% to 24% of the crashes and near crashes observed (Klauer et al., 2006). A study of male drivers in the United Kingdom, found that fatigue was a contributing factor to approximately 10% of crashes (Maycock, 1997).

Despite these consequences, many individuals continue to drive while drowsy. A National Sleep Foundation study found that 28% of drivers drive while drowsy at least one day per month, and 28% of those individuals reported that they had actually fallen fully asleep while driving (National Sleep Foundation, 2009). A focus group analysis found that commonly accepted remedies such as rolling down the windows, turning on the air conditioning, and consuming caffeine are neither robust nor reliable. The report also indicated that drivers were enthusiastic about new technologies that could alert them before they fall asleep (Nelson et al., 2001). Together these findings suggest a need for new technologies to detect drowsiness before serious mishaps occur. This work briefly explores current methods for drowsiness detection and presents a new method which is unobtrusive, inexpensive to implement, and performs comparably with current approaches.

Current methods for drowsiness detection

A wide range of techniques for drowsiness detection have been explored in recent years (Dong et al., 2011). These algorithms have typically differed across several dimensions including: the number and type of predictors, definition of drowsiness, timeliness, and classification algorithm type.

The predictor variables in most algorithms are typically driver-based physiological measures, or vehicle-based behavioral measures, although environmental measures are occasionally included. Physiological measures have included: brain activity (Dinges et al., 1998), eye closure (Dinges et al., 1998), facial expressions (Ji et al., 2004), and head position (Dinges et al., 1998). Vehicle-based measures typically involve lateral control, such as steering-wheel angle (Eskandarian & Mortazavi, 2007; Krajewski et al., 2009; Sayed & Eskandarian, 2001), and lane position deviation (Hanowski et al., 2008). Environmental variables are significantly less common than the other two categories; they include road curvature, and demand metrics (Sayed & Eskandarian, 2001). These variables are typically included in conjunction with driver-based measures to explain variance caused by actions such as planned turns and stops.

Currently, there is no gold standard for an operational definition of drowsiness while driving. This is due to large individual differences between drivers and the fact that drowsiness is both a chronic and an acute phenomenon (Grace & Stewart, 2001). This deficit has led the research community to define drowsiness with a wide range of dependent measures in drowsiness detection algorithms. These measures are often based on experimental condition (day vs. night driving, or sleep deprived vs. awake), subjective ratings, such as the Stanford Sleepiness Scale (Hoddes, 1973), or the presence of simulated accidents or lane departures.

The measures that define and predict drowsiness often determine the "timeliness" of an algorithm. Timeliness can be defined as the algorithm's ability to detect acute drowsiness episodes. For instance, an algorithm that detects a drowsiness-related lane departure with steering input aggregated over the preceding minute would have high timeliness, whereas an algorithm designed to detect aggregate subjective ratings of sleepiness with the cumulative time that the driver has been awake would have low timeliness. In the latter case, information must be aggregated over a large time window, making detection at the timescale of seconds infeasible.

Many classification algorithms have been used to detect drowsiness and other forms of impairment. Algorithms with a single independent variable typically employ a single threshold to separate drowsy and non-drowsy drivers. More complicated algorithms have used support vector machines (Lee et al., 2010; Liang, Reyes, et al., 2007), Bayesian networks (Liang, Lee, et al., 2007), artificial neural networks (Eskandarian & Mortazavi, 2007; Sayed & Eskandarian, 2001), and decision trees (Lee et al., 2010).

The most promising and popular algorithm for drowsiness detection is PERCLOS. PERCLOS is a measure of percent closure of the eyes, averaged over a brief time window, often as short as one minute (Wierwille et al., 1994). Several studies have evaluated PERCLOS and found that it can achieve accuracies of over 90% in simulator studies (Wierwille et al., 1994) and real world conditions (Tijerina et al., 1999). The success of PERCLOS and its wide acceptance makes it an excellent algorithm for comparison with new methods.

METHODS

The intent of this study is to design and evaluate an effective algorithm for drowsiness detection that is efficient, unobtrusive, and easily applicable with current technology. This section discusses the design process of this new algorithm beginning with a description of the data collection process that supported and informed its development.

Drowsy driver data

The data used in this analysis were collected in a study at the University of Iowa's National Advanced Driving Simulator (NADS). The study was a 2 x 2 x 3 x 3 mixed design, with drive session order, age, and gender as between-subjects variables, and drowsiness as a within-subjects variable. The details of this experiment are briefly described in this section. A thorough explanation can be found in Brown et al. (2011).

Participants and task. Seventy two individuals participated in three drives of approximately 30 minutes each. All drivers were healthy males and females from one of three age groups, 21-34, 38-51, or 55-68 years of age. They were required to be licensed drivers who had driven at least 10,000 miles per year for each of the last two years. Each drive consisted of urban, highway, and rural segments. The drives occurred during separate segments of the day, with one drive occurring during the daytime, one in the early evening, and one late in the evening (and early morning). These times were designed to induce significant variance in drowsiness. The evening drives were blocked into one night and separated by a period of three days from the daytime drive. The order of these drives was counterbalanced across subjects. In addition to the two data collection sessions, drivers participated in a screening visit where they were screened for drug and alcohol use, trained on simulator operation, and tested for simulator sickness. Participants were compensated $250 for their full participation or a prorated amount for less than full participation.

Data collection. Data were collected at 60Hz. Two separate sources of data were collected from the simulator and a Face Lab 5.0™ (Seeing Machines, Canberra, Australia) eye tracking system which was mounted on the dashboard above the steering wheel. A separate data set was created from the raw data streams for algorithm evaluation. This "evaluation data set" affords clear definitions of drowsiness and control over random variance associated with subjects and road conditions.

Algorithm evaluation data set. The data from the evaluation data set were limited to a 60s window before drowsy-related lane departures or 60s matched cases and sampled at reduced frequency of 1Hz.

Drowsy-related lane departures were defined through a thorough video analysis of all departures detected in the raw simulator data. Each case was evaluated by two reviewers using a modified Observer Rating of Drowsiness (ORD) scale. The ORD scale is a continuous rating between 0 and 100, based on 60s of analysis leading up to a lane departure (Wierwille et al., 1994). A modified ORD (mORD) was used in this study that binned ratings into five categories ranging from Not Drowsy (mORD = 1 or ORD < 12) to Extremely Drowsy (mORD = 5 or ORD > 90). Departures with mORD ratings of greater than 2 were classified as drowsy. Matched cases were selected from only verifiably awake drivers. These drivers were identified through their performance on a modified Psychomotor Vigilance Task (Wilkinson & Houghton, 1982) and their subjective ratings of their own drowsiness on two scales: the Stanford Sleepiness Scale (Hoddes, 1973) and a retrospective sleepiness scale. Only participants who were classified as verifiably awake on all three measures in the daytime drive were considered for the matched data. From this subset, matched cases were selected to either match both driver and road segment with a drowsy instance, or only match road segment with a drowsy instance.

The final evaluation data set consisted of 578 total instances with 162 drowsy instances, 80 awake cases matched to driver and road segment, and 336 awake cases matched to only road segment.

Algorithm design

The goal of this algorithm design process was to create an algorithm that performs favorably with respect to the current standard using only data available in all vehicles. The goals of minimizing processing time and data manipulation were also considered.

Variable selection. Within these constraints, Steering wheel angle (steering-angle) was selected as the single metric of analysis. Steering-angle has been included in other combination algorithms (Krajewski et al., 2009) and has been successfully used by itself but only after significant data preprocessing (Eskandarian & Mortazavi, 2007; Sayed & Eskandarian, 2001). The goal of limiting processing time makes such pre-processing undesirable for this application. In contrast to previous work, our proposed algorithm uses unprocessed steering data. The data are organized by steering angle at each second from 60s before a lane departure (or matched case) to 6s before the lane departure. This 6s buffer was selected to ensure that there is enough time in between detection and departure for drivers to perceive and react to a warning. Table 1 shows a sample of model input data.

Table 1 Example of data input to the steering-angle model. Note that all measurements are in degrees.

Drowsy	60s	59s	58s	…
0	-0.9946	-0.5837	-0.9308	…
1	-0.4116	-0.1926	-0.9264	…

Classifier selection. One limitation on the use of unfiltered human behavioral data is the presence of noise. Additionally human behavior is rarely linearly separable—into drowsy and non-drowsy drivers in this case. Given these restrictions, it was important to select a classifier that powerfully separates nonlinear data and is robust to noise. Another consideration is that the organization of the data causes successive features to be highly dependent on one another. Therefore it was important to avoid algorithms that strongly assume independence of features. The random forest (RF) algorithm developed by Breiman (2001) fits these criteria.

The random forest algorithm's training output is a series of decision trees with randomly selected features. Classification is performed by a majority vote of the predictions from the trees for each new instance. The training algorithm consists of two steps:

1. Draw *n* bootstrap samples from the training data set
2. For each sample train a decision tree using a random sample of *m* features to determine the best split amongst the data.

The forest can be optimized over both *n* and *m* (Liaw & Weiner, 2002).

Algorithm Training. The relatively small number of samples in the final data set makes partitioning the data into training and test sets impractical. On the other hand, the specialization of this data set increases the likelihood of overfitting. A ten-fold cross validation was used for all algorithms to limit the problem of overfitting. All model training and analyses were completed using the caret (Kuhn, 2008; Kuhn et al., 2011) package in R (R Development Core Team, 2010). The caret package implements random forests through the randomForest package (Liaw & Weiner, 2002). The graphical and statistical analyses were also conducted in R using the pROC package (Robin et al., 2011).

RESULTS

The data from the ten-fold cross validation were used to calculate the accuracy, Positive Predictive Value (PPV), and the Area under the Receiver Operating Characteristics (ROC) Curve (AUC) for the steering model. AUC (a non-parametric version of d') is a robust metric for analyzing binary classifiers because it is insensitive to proportions of positive and negative input (Fawcett, 2004). The PERCLOS model was analyzed using a fixed threshold value of 2.96 (% eye closure/ min), which corresponded to the threshold of maximum AUC from the ROC analysis. An estimation of standard deviations for the accuracy and PPV of PERCLOS were calculated by bootstrapping with 500 replicates.

A summary of accuracy and PPV results is presented in Table 2. Figures 1 and 2 show the ROC curves with bootstrapped (500 replicates) standard error bands. The horizontal axis of these figures indicates the false alarm (1-specificity) rate and the vertical axis indicates the hit rate (sensitivity).

Table 2 Mean accuracy and PPV values for PERCLOS and the steering-model with standard deviation in parentheses.

Model	Accuracy	PPV
PERCLOS	0.55 (0.06)	0.88 (0.08)
RF-steering-model	0.79 (0.04)	0.80 (0.02)

Table 2 shows that the mean accuracy of the RF-steering-model was 24% higher than PERCLOS, a 43.6% improvement. A Fisher's Exact Test (Fisher, 1922) showed that the difference was statistically significant ($p < 0.001$). PERCLOS did have a higher mean PPV; however a second Fisher's Exact Test showed that this difference was not significant, ($p = 0.18$).

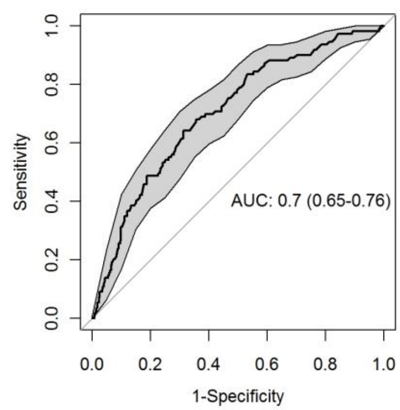

Figure 1: ROC curve for the RF-steering-model, including the AUC with a bootstrapped confidence interval. The grey diagonal line represents a random classification.

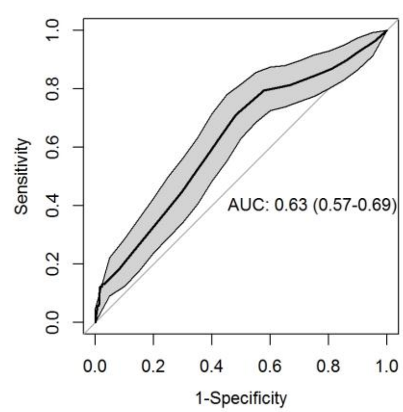

Figure 2 ROC curve for PERCLOS, including the AUC with a bootstrapped confidence interval. The grey diagonal line represents a random classification.

Figure 1 and Figure 2 show that the RF-steering-model has a higher AUC than PERCLOS. Additionally the AUC of the PERCLOS curve is only slightly better than random in the lower quadrant of the graph. This indicates that the model will only achieve high detection rates when the threshold is lowered. This result agrees with the disparity between accuracy and PPV for PERCLOS in Table 2. There is no established statistical test for significance between AUC values; however several methods have been proposed to analyze such differences. The pROC package in R employs two methods: a non-parametric method proposed by DeLong et al. (1988), and a bootstrapping method. The DeLong method uses generalized U-statistics to estimate a covariance matrix for the ROC curves and compares multiple curves using contrasts. The bootstrapping method divides the difference in AUC of the original curves by a standard deviation of the difference in AUC derived from n stratified samples of the original data, and compares that value (D) to 0 using a z-test. The results of the DeLong method showed that the difference in AUC between the RF-steering-model and PERCLOS was significant: D(1028.29) = -2.01, p = 0.045. The bootstrapping method with 2000 replicates also showed a significant difference: D = 2.03, p = 0.425.

A brief descriptive analysis of the quality of eye-closure measurements was also conducted. The analysis was motivated by the large difference between the PERCLOS results presented and previous estimates of accuracy. This analysis showed that the mean eye-closure measurement confidence was 0.52 (SD = 0.16) for correct predictions and 0.54 (SD = 0.14) for incorrect predictions. These results suggest the presence of drowsiness did not affect eye-tracker confidence however the relatively low level of confidence in general indicates a need for further analysis. As a follow-up to these results a new PERCLOS model was fit to a data set that only included cases where the eye-closure measurement confidence in both eyes was greater than 0.6. The data set included 32 drowsy instances and 58 awake instances. The AUC of this high confidence model was still 0.67, which suggests that the model fit is not highly sensitive to changes in confidence.

In addition to PERCLOS the RF-steering-model was compared against two other models using only steering data. The first model used a decision tree algorithm trained on steering-angle distribution parameters. The parameters included: mean, standard deviation, skewness, and kurtosis. The second model used only steering reversals. These two models were selected because they are simpler to implement and explain compared to the random forest model. Additionally they represent variables and methodologies that have been previously applied to drowsiness detection.

Figure 3 shows the comparison between the three models on an ROC plot. Further analysis using the DeLong method indicated that the RF-steering model had a significantly higher AUC than both the distribution parameter model (D(1270.6) = -1.94, p = 0.05) and the steering-reversal model (D(1092.6) = -5.49, p < 0.001). Variable importance measures of the distribution parameter model indicated that the standard deviation was the most important variable used in classification, and that the mean, skewness, and kurtosis had little predictive power (less than 10% importance relative to standard deviation).

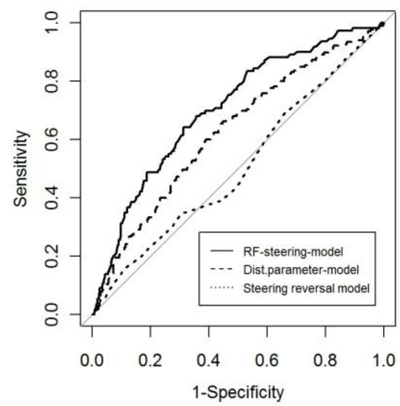

Figure 3: ROC curves for the RF-steering-model, the distribution parameter model, and the steering reversal model.

DISCUSSION

The purpose of this research was to explore and develop timely models for predicting drowsiness-related lane departures which use input from unobtrusive sensors, perform comparably to the current standard, and could be easily implemented in current vehicles.

The proposed model uses a 60s moving window of steering wheel angle input data sampled at 1Hz to predict the presence of a drowsiness-related lane departure 6s in advance of its occurrence. The results of this model are promising particularly when compared to PERCLOS, the current standard for drowsiness detection, and simpler steering models. The RF-steering-model and PERCLOS have comparable Positive Predictive Value; and the RF-steering-model had significantly higher accuracy. The AUC of the RF-steering model was higher than the AUC of PERCLOS, and the difference was statistically significant using both the DeLong and Bootstrapping methods. The RF-steering model also had a significantly higher AUC than simpler models using distribution parameters and steering reversals. These results agree with previous conclusions that provide evidence that steering wheel angle is a powerful predictor of drowsiness (Eskandarian & Mortazavi, 2007; Krajewski et al., 2009; Sayed & Eskandarian, 2001). Additionally they demonstrate that the steering wheel angle is particularly powerful for predicting acute effects of drowsiness such as lane departures.

These results do not suggest that PERCLOS should be discarded or ignored in the future. They merely suggest that steering angle is a more robust metric for drowsiness detection on this time scale. . In the future it may be particularly advantageous to employ a combination of the RF-steering model and PERCLOS in a hierarchical structure. This structure could use information on various time scales such as cumulative time awake, PERCLOS, and steering-angle to predict drowsiness at both long and short time horizons.

It is important to acknowledge that this model is somewhat limited by its frequency of “false positive” classifications, 0.39 (SD = 0.09) at the threshold that produces maximum AUC. The false positive rate limits the type and severity of interventions this model could support. However, these limitations do not eliminate the possibility of pairing the model and an intervention that could significantly reduce crashes and near crashes caused by drowsiness. Future work in this area will consider the nature of the random forest algorithm, and the features of the steering data to which it is particularly sensitive, with the ultimate goal of reducing the false positive rate.

Acknowledgements

The data presented in this paper were collected as part of a project sponsored by the National Highway Traffic Safety Administration (NHTSA) and other parts of this study were published as a NHTSA technical report, under the leadership of Eric Traube. The authors would also like to acknowledge the assistance provided by Eric Nadler.

REFERENCES

Breiman, L. (2001). Random Forests. *Machine Learning, 45*, 5-32.

Brown, T., Lee, J. D., Schwarz, C., Fiorentino, D., & McDonald, A. (2011). *Final Report: Advanced Countermeasures for Multiple Impairments*. Washington DC.

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the Areas Under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics, 44*, 837-845.

Dinges, D. F., Mallis, M. M., Maislin, G., & Powell IV, J. W. (1998). *Final Report: Evaluation of Techniques for Ocular Measurement as an Index of Fatigue and as the Basis for Alertness Management*. Washington DC.

Dong, Y., Hu, Z., Uchimura, K., & Murayama, N. (2011). Driver Inattention Monitoring System for Intelligent Vehicles: A Review. *IEEE transactions on intelligent transportation systems, 12*(2), 596-614.

Eskandarian, A., & Mortazavi, A. (2007, June 13-15, 2007). *Evaluation of a Smart Algorithm for Commercial Vehicle Driver Drowsiness Detection.* Paper presented at the 2007 IEEE Intelligent Vehicles Symposium, Istanbul, Turkey.

Fawcett. (2004). ROC Graphs: Notes and Practical Consideration for Researchers. *ReCALL, 31*(HPL-2003-4), 1-38.

Fisher, R. A. (1922). On the interpretation of Ξ^2 from Contingency Tables, and the Calculation of P. *Journal of the Royal Statistical Society, 85*(1), 87-94.

Grace, R., & Stewart, S. (2001, August 14-17, 2001). *The Co-Pilot: A Low-Cost Drowsy Driver and Driver Inattention Monitor.* Paper presented at the The International Driving Symposium on Human Factors in Driver Assessment, Training And Vehicle Design, Aspen, CO.

Hanowski, R. J., Bowman, D., Alden, A., Wierwille, W. W., & Carroll, R. (2008). *PERCLOS+: Development of a Robust Field Measure of Driver Drowsiness*. Paper presented at the 15th World Congress on Intelligent Transport Systems and ITS America's 2008 Annual Meeting, New York.

Hoddes, E. (1973). Quantification of Sleepiness: A New Approach. *Psychophysiology, 10*(4), 431-436.

Ji, Q., Zhu, Z., & Lan, P. (2004). Real-Time Nonintrusive Monitoring and Prediction of Driver Fatigue. *IEEE transactions on vehicular technology, 54*(4).

Klauer, S. G., Dingus, T. A., Neale, V. L., Sudweeks, J., & Ramsey, D. J. (2006). *The impact of driver inattention on near-crash/crash risk: An analysis using the 100-car naturalistic driving study data*. Washington D.C.

Krajewski, J., Golz, M., & Sommer, D. (2009). *Detecting Sleepy Drivers by Pattern Recognition based Analysis of Steering Wheel Behaviour*. Paper presented at the Der Mensch im Mittelpunkt technischer Systeme, Dusseldorf.

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software, 28*(5).

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., & Engelhardt, A. (2011). caret: Classification and Regression Training (Version R. package version 5.01-001). Retrieved from http://CRAN.R-project.org/package=caret

Lee, J. D., Fiorentino, D., Reyes, M. L., Brown, T., Ahmad, O., Fell, J., . . . Dufour, R. (2010). *Assessing the Feasibility of Vehicle-Based Sensors to Detect Alcohol Impairment* (DOT HS 811 358). Washington DC.

Liang, Y., Lee, J. D., & Reyes, M. L. (2007). Non-intrusive detection of driver cognitive distraction in real-time using Bayesian networks. *Transportation Research Record: Journal of the Transportation Research Board, 2018*, 1-8.

Liang, Y., Reyes, M. L., & Lee, J. D. (2007). Real-time detection of driver cognitive distraction using Support Vector Machines. *IEEE Transactions on Intelligent Transportation Systems, 8*(2), 340-350.

Liaw, A., & Weiner, M. (2002, December 2002). Classification and regression by randomForest, *R News*, pp. 18-22. Retrieved from http://cran.r-project.org/doc/Rnews/Rnews_2002-3.pdf

Maycock. (1997). Sleepiness and driving: The experience of U.K. car drivers. *Accident Analysis & Prevention, 29*(4), 453-462.

National Highway Traffic Safety Administration. (2011). Research on Drowsy Driving Retrieved 21 December, 2011, from http://www.nhtsa.gov/Driving+Safety/Distracted+Driving/Research+on+Drowsy+Driving

National Sleep Foundation. (2009). *2009 Sleep In America™ Poll: Summary Of Findings*. Washington DC.

Nelson, T. F., Isaac, N. E., & Graham, J. D. (2001). *Development and Testing of Countermeasures for Fatigue Related Highway Crashes*. Washington DC.

R Development Core Team. (2010). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J., & Muller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves (Version 22): BMC Bioinformatics. Retrieved from http://www.biomedcentral.com/1471-2105/12/77/

Sayed, R., & Eskandarian, A. (2001). *Unobtrusive drowsiness detection by neural network learning of driver steering*. Paper presented at the Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering.

Tijerina, L., Gleckler, M., Stolzfus, M., Johnston, S., Goodman, M. J., & Wierwille, W. (1999). *A Preliminary Assessment of Algorithms for Drowsy and Inattentive Driver Detection on the Road*. Washington DC.

Wierwille, W., Ellsworth, L., Wreggit, S., Fairbanks, R., & Kim, C. (1994). *Final Report: Research on vehicle-based driver status/performance monitoring: development, validation, and refinement of algorithms for detection of driver drowsiness*. Washington DC.

Wilkinson, R. T., & Houghton, D. (1982). Field test of arousal: A portable reaction timer with data storage. *Human Factors, 24*, 487-493.