U.S. Department of Transportation
National Highway Traffic Safety Administration
TRAFFIC SAFETY FACTS
Research Note
NHTSA
DOT HS 813 003 October 2020

Quantifying Drowsy Driving

Drowsy driving remains a problem for motorists. According to the National Highway Traffic Safety Administration's National Center for Statistics and Analysis (NCSA), drowsy driving was a contributing factor in 775 deaths in 2018, or 2.1% of total fatalities involving motor vehicle crashes on U.S. roadways (NCSA, 2019) and, in 2015, an estimated 1.4% of all police-reported crashes (NCSA, 2017). However, these statistics, which are based upon information in police crash reports, likely underestimate the extent of the drowsy driving problem. Examining NHTSA's National Automotive Sampling System Crashworthiness Data System, the American Automobile Association Foundation for Traffic Safety (AAAFTS) used multiple imputation to estimate the percentage of fatal crashes that involved a drowsy driver (Tefft, 2014). AAAFTS estimated that 21% of fatal crashes involved a drowsy driver.

The current project explored the feasibility of using machine learning algorithms to identify drowsy driving episodes in large-scale naturalistic driving datasets. The Strategic Highway Research Program 2 (SHRP2) Naturalistic Driving Study (NDS) collected data from more than 3,400 participating drivers in six States, yielding time series data for more than 5 million trips (Virginia Tech Transportation Institute, 2020). The SHRP2 NDS database contains vehicle variables and video data, including a view of the driver's face. Given the combination of driver-facing camera views and extensive time history vehicle data such as acceleration and lane deviation, it may be possible to identify when drowsiness was involved in crashes and near-crashes as well as in incident-free trips. To that end, the current research developed two machine learning algorithms to identify possible drowsy driving episodes within the SHRP2 NDS database. For the purposes of this project, an algorithm is defined as a process or set of rules to be followed in calculations or other problem-solving operations. The first algorithm used time history vehicle data that measured driving behaviors associated with drowsy driving, such as lane deviations. The second algorithm used face video data to identify driver's behaviors associated with drowsiness, such as head nodding.

Observer Rating of Drowsiness

Driver impairments in all trip segments ("epochs") involving a crash or near-crash, as well as a sample of baseline driving, were manually coded. One category of possible driver impairments is "Drowsy, sleepy, asleep, fatigued," and the category is defined as "Subject vehicle driver exhibits obvious signs of being asleep or tired, or is actually asleep while driving, degrading performance of the driving task" (VTTI, 2015). The research team obtained 589 one-minute epochs with drowsiness noted as a driver impairment and 200 additional baseline epochs in which there was no apparent driver impairment due to drowsiness. To determine drowsiness level, three separate analysts coded the 789 epoch videos using the Observer Rating of Drowsiness (ORD) protocol (Wierwille & Ellsworth, 1994; Wiegand et al., 2009). The ORD protocol involves evaluating on a continuous scale the driver's gestures, facial tone, and behavior for drowsiness and provides a reliable instrument for identifying drowsiness (Wiegand et al., 2009). Due to poor video quality or missing data, researchers could not code some epochs with the ORD protocol. Ultimately, 741 (94%) of the epochs were retained and rated. This project used the average of the three ORD ratings as "ground truth" for testing the two drowsy driving identification algorithms (i.e., the algorithms would be valid if they could successfully identify drowsy driving episodes initially coded using the ORD protocol).

Algorithm 1: Time History Vehicle Data And ORD

The first algorithm used time history vehicle data, including vehicle-based sensor data, to identify instances of drowsy driving. Researchers removed long segments of missing data, imputed short sections of missing data using truncation and interpolation, removed turns from the data as they can introduce a combination of intentional lane departure and slow speeds that are different from normal driving behaviors, and truncated trips whenever the speed was less than 30 kph to remove low-speed maneuvers such as stopping at intersections. Researchers then divided each ORD-coded epoch into 6-second subsequences for analysis purposes. For each subsequence, the researchers calculated the 25th percentile, median, 75th percentile, and standard deviation for the following sensor-based measures: yaw rate (or the rotation of the car around the z axis) and lane position as a deviation from the center of the lane. Researchers also calculated lane position as a slope relative to the lane center, i.e., they examined the change in distance from the center line over time. Researchers then calculated the 25th percentile and 75th percentile for the epochs across all the subsequences for yaw rate, lane position as a deviation from the center of the lane, and lane slope (see Table 1). Researchers also derived the number of lane departures and lane changes in the epoch. If a single lane departure was categorized as left at the beginning and right at the end, or vice versa, then it was classified as a lane change because a driver crossed the center line. This data enabled the calculation of the number of lane departures with subsequent corrections. Additional time history vehicle data included time the trip began (in binned 3-hour windows) and trip duration (i.e., the time from the start of the trip to the start of the epoch). Researchers used these measures as the input features for algorithm development.

Table 1
Algorithm Inputs

Yaw rate	
1	25th percentile of all subsequences of gyro_z_25%. Gyro_z_25% is the 25th percentile of yaw rate, the rate of change around the yaw axis (or the rotation of the car around the z axis).
2	25th percentile of all subsequences of gyro_z_75%. Gyro_z_75% is the 75th percentile of yaw rate, the rate of change around the yaw axis.
3	75th percentile of all subsequences of gyro_z_25%. Gyro_z_25% is the 25th percentile of yaw rate, the rate of change around the yaw axis.
4	75th percentile of all subsequences of gyro_z_75%. Gyro_z_75% is the 75th percentile of yaw rate, the rate of change around the yaw axis.
Lane Position (deviation from center of lane)	
5	25th percentile of all subsequences of lanepos_25%. Lanepos_25% is the 25th percentile of lane position.
6	25th percentile of all subsequences of lanepos_75%. Lanepos_75% is the 75th percentile of lane position.
7	75th percentile of all subsequences of lanepos_25%. Lanepos_25% is the 25th percentile of lane position.
8	75th percentile of all subsequences of lanepos_75%. Lanepos_75% is the 75th percentile of lane position.
Lane Departures	
9	The number of lane departures.
Slope of Lane Position	
10	25th percentile across all subsequences of the slope of lane position. Slope of lane position is a measure of the change in lane position across a subsequence.
11	75th percentile across all subsequences of the slope of lane position. Slope of lane position is a measure of the change in lane position across a subsequence.
Trip Start Time and Duration	
12	The 3-hour time bin at the start of the trip.
13	Trip duration at the start of epoch.

Researchers then created a vector for each epoch using the 13 features described above. These vectors served as input to several machine learning algorithms that classified epochs as drowsy or not. After eliminating epochs with significant missing time history vehicle data, researchers retained 432 with complete vectors. The ORD scale ranged from 0 to 100, with 0 considered to be not drowsy, 50 to 69 considered to be moderately drowsy, and 70 or higher considered to be very drowsy. Researchers used two different thresholds to divide drivers' ORD scores into drowsy and non-drowsy: 50 and 70.

Researchers applied four tree-based classifiers from the Python library "scikit-learn" to the data: gradient boosting classifier, random forest classifier, extra trees classifier, and ada boost classifier. The scikit-learn library can return the relative importance of each input feature to the training of a tree-based classifier. The relative importance of each feature can be compared.

ORD Threshold = 50

Researchers reserved one third (33%) of the epochs as a test set to evaluate the performance of the algorithm.

Of the four tree-based learners described, the extra trees classifier performed the best at classifying drowsy driving episodes and was selected to be the algorithm for an ORD threshold of 50.

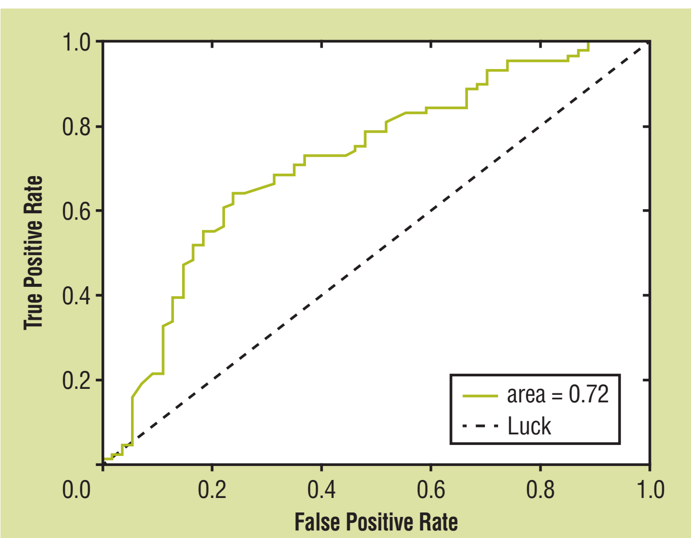
When applied to the test set, the researchers obtained the confusion matrix shown in Table 2. The confusion matrix describes the extent that the algorithm correctly identified the epochs in which the driver was drowsy (per the ORD ratings). Thus, for an ORD threshold of 50, epochs with ORD ratings equal to or greater than 50 were considered "drowsy." The Matthews Correlation Coefficient (MCC), a measure of the quality of binary classifications, was 0.25. Like other correlation coefficients, a value of 0 corresponds to random, or no correlation, 1 is complete agreement, and -1 is complete disagreement. A correlation of 0.25 represents modest agreement.

The receiver operating characteristic (ROC) curve for the algorithm is shown in Figure 1, with its area under the curve (AUC) marked on the graph. An AUC of 1.0 would perfectly identify drowsiness and an AUC of 0.5 (Figure 1, dotted line) would identify drowsiness at the level of chance, or the accuracy if the algorithm was guessing randomly about an epoch's classification. The AUC corresponding to the ORD threshold of 50 was 0.72.

Table 2
Confusion Matrix With ORD Threshold of 50

		Algorithm Classification	
		False	True
ORD Rating	False	16	38
	True	9	80

Figure 1
ROC Curve With ORD Threshold of 50



Researchers rank ordered the features in terms of importance of generating the model's predictions. For classifying drowsy epochs at an ORD threshold of 50 (moderately drowsy), trip start time (12) was the most important feature (while trip duration (13) was the second most important, its importance was similar to many lane position measures).

ORD Threshold = 70

In this case, epochs with ORD ratings of 70 or greater were considered "drowsy," indicating a more extreme level of drowsiness. Of the four tree-based learners described, the gradient boosting classifier performed the best at classifying drowsy driving episodes and was selected to be the algorithm for an ORD threshold of 70.

When applied to the test set, the researchers obtained the confusion matrix shown in Table 3. The MCC was calculated as 0.33, a modest correlation. The ROC curve with marked area (AUC) is shown in Figure 2. The AUC for an ORD threshold of 70 was 0.76.

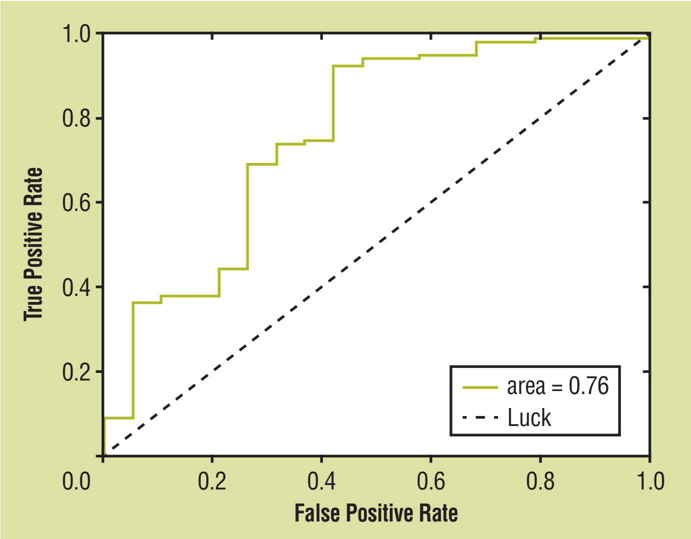
Table 3
Confusion Matrix With ORD Threshold of 70

		Algorithm Classification	
		False	True
ORD Rating	False	6	13
	True	6	118

The importance of each feature for an ORD score of 70, or very drowsy, was rank ordered. The two most important features were the 75th percentile of the lane position slope (11) and the 25th percentile of the lane position slope (10), variables that describe the amount of lane deviation over an epoch. The important features from the threshold of 50, trip start time (12) and trip duration (13), were the fifth and sixth most important for the threshold of 70.

The results indicate moderate success for the classifiers used for identifying cases of moderately drowsy driving (ORD of 50) or of severely drowsy driving (ORD of 70). Three features are in the top five for both thresholds when rank ordered: trip duration (13), 25th percentile of all subsequences of the 25th percentile of lane position (5), and 25th percentile of all subsequences of the 75th percentile of lane position (6). Thus, information about the length of time a driver had been driving and the deviation from the center of the lane were generally useful for identifying drowsy epochs. On the other hand, information about the slope of the lane position (10, 11) became the most important features after raising the

Figure 2
ROC Curve With ORD Threshold of 70



threshold to 70, indicating that lane position was useful for classifying "very drowsy" epochs. The most important feature for the moderately drowsy classifier was the trip start time, followed by the trip duration, indicating that information about time of day and length of trip were useful for classifying "moderately drowsy" epochs. These results indicate that moderate drowsiness could be suggested by typical predictors of time of day and time on task, whereas very drowsy drivers have more pronounced drifts in the lane that can be detected using vehicle sensors. These factors may be used to narrow the search for drowsy driving episodes within naturalistic driving data but may not be robust enough to correctly predict all or most episodes of drowsy driving.

Algorithm 2: Face Video and ORD Ratings

To complement the time history vehicle data as a means of detecting episodes of drowsy driving, the project team also examined the driver's face video, including the minute of the epoch with an ORD rating and one minute prior. Researchers developed a machine learning algorithm to detect epochs that had ORD ratings above 50, or moderately drowsy. First, machine vision algorithms processed the face video to extract the location of facial landmarks (e.g., eyes and nose) and, from those, the orientation of the drivers' head (e.g., yaw, pitch, and roll of the head) using video analysis software (Smith, Dyer, Chitturi, & Lee, 2017). In addition, researchers produced landmark coordinates for the eyes and mouth to determine eye/mouth openness. After extracting the face data for a frame of the video, researchers used the data as a starting point for the analysis of the next frame. In the final stage, the research team initialized the next frame using the estimate of head position and orientation from the previous video frame.

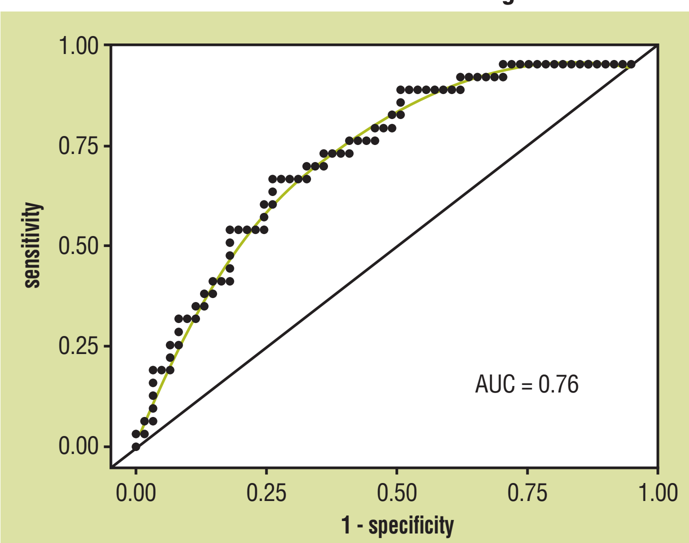
The data then went through a three-stage process to develop the model for predicting drowsy driving. The first stage highlighted and filtered highly uncertain or inconsistent data. Because video quality varied considerably across epochs, not all data extracted by the machine vision algorithm accurately represented the drivers' facial features and head orientation. For this reason, researchers filtered the output of the machine vision algorithm to eliminate epochs where the machine vision algorithm produced low confidence scores. Low confidence typically reflected poor lighting conditions, such as nighttime driving or extremely bright situations. After this filtering process, a total of 594 epochs remained. These epochs were split into training (85%) and testing (15%) sets. The epochs were a stratified random sample that created an equal number of drowsy and alert cases (according to ORD ratings).

The second stage involved deriving features from the output of the machine vision algorithm, i.e., summary statistics that are plausibly related to drowsiness. Examples included PERCLOS, the percent of time the eyes are closed over a 60-second window, as well as measures of the changes of head orientation that could indicate drowsiness-related head bobs. Beyond these features, researchers also included the median and interquartile range of the head position and its yaw, pitch, and roll, as well as eye and mouth position. Additional features were developed by combining different factors (such as head position with size of mouth gap).

The third stage fit a machine learning algorithm to the features to predict drowsiness. Many possible algorithms can serve this purpose, and researchers implemented five: general linear model, lasso, support vector machine, random forest, and extreme gradient boosting. Researchers trained each of these five models on the facial features, using a 10-fold cross-validation tuning process. If a model required its own tuning parameters, such as the number of decision trees in the random forest, researchers estimated these by fitting the model with a range of these parameters and then selecting the one that produced the best performance. The research team repeated the 10-fold cross-validation process twice, with each repetition selecting different training data. The repeated cross-validation provides an estimate of how well the model performs with data that were not included in the training set.

Figure 3 summarizes the performance of the models on the training and testing data. The area under the ROC curve measure (AUC) on the y-axis represents the overall accuracy of the model. It indicates how the model separates instances of drowsy and alert driving. As stated above, an AUC of 1.0 would perfectly identify drowsiness, and an AUC of 0.5 would identify drowsiness at the level of chance. The extreme gradient boosting algorithm had the best performance on the test data (AUC = 0.76) and the smallest variation of performance with the cross-validation on the training data. The x-axis represents the rate of false alarms at different thresholds and the y-axis represents the true positive rate. The diagonal line represents chance performance. The model performance is imperfect, but well above chance. The most important features for predicting drowsiness included factors such as eye gap, mouth gap, pitch, yaw, and roll of the head.

Figure 3
ROC Plot for the Extreme Gradient Boosting Model



Exploring Misclassifications

Although the time history vehicle data algorithm and the machine-vision face video algorithm performed better than chance at predicting drowsy driving epochs, they still made a fair number of misclassifications. Upon further review, patterns emerged that helped to explain the misclassifications. When the ORD raters classified the driver as drowsy, but the time history vehicle data algorithm classified the driver as not drowsy, there tended to be yawning and other clear indicators of drowsiness on the video that could not be captured by the time history vehicle data model. For some epochs, "swerving" unrelated to drowsiness was observed, which may have caused the time history vehicle data algorithm to classify the epoch as drowsy when the ORD rating was not drowsy. In addition, for some videos, the coders may have been influenced toward higher ORD ratings due to unusual facial features (like droopy eyes or squinting) or blurry video. Finally, when there was disagreement between the ORD and the machine-vision face algorithm, there were more misses than false alarms, and about half of the misses occurred when the driver was engaged in some other activity (talking to someone, looking at something in the car) and also appeared drowsy.

To better understand the nature of the misclassifications, researchers used Fisher's exact test (two-tailed) to determine if the distribution of epoch characteristics differed between the outcome categories of the video epochs reviewed. Driver factors (epoch characteristics) included whether the driver was talking to someone, squinting, if she/he had unusual facial features, if the driver was looking at something in the car, was distracted, or had displayed any other unusual behavior. The distribution of counts across the two groups (one or more driver factors noted vs. no driver factors noted) varied significantly between outcome categories (p = .013). The highest percentage of epochs with driver factors noted (62%) occurred for the outcome category where there was disagreement between ORD and the machine vision face algorithm. It is possible that the face algorithm sometimes had trouble with the correct interpretation of unusual facial features or facial behaviors.

Researchers also classified traffic conditions as heavy, medium, or light for each epoch. The distribution of counts across the two groups (heavy or medium vs. light traffic conditions) varied significantly between outcome categories (p = .0004). In cases where the time history vehicle data algorithm predicted the driver to be drowsy, but the ORD rating did not, 73% of the epochs occurred in heavy or medium traffic. It is possible that in these cases, the algorithm picked up instances of hard braking or lane departures that were in response to the traffic rather than the driver's physical state. Similarly, for cases where the time history vehicle data algorithm classification was uncertain, 63% of the epochs occurred in heavy or medium traffic.

For each epoch reviewed, researchers noted unusual driving behaviors such as swerving, sudden braking, tailgating, maintaining an especially long headway, or other potentially unsafe driving behaviors. The distribution of counts across the two groups (one or more unusual driving behaviors noted vs. no unusual driving behaviors noted) differed significantly by outcome category (p = .005). In most epochs reviewed, researchers noted one or more unusual driving behaviors. For both outcome categories where the time history vehicle data algorithm classification was drowsy driving (i.e., the ORD rating also classified the driver as drowsy or the ORD rating did not classify the driver as drowsy, but the algorithm classified the driver as drowsy), the epochs contained at least one unusual driving behavior. By contrast, among the cases that the time history vehicle data algorithm classified as not drowsy, 75% of epochs contained one or more unusual driving behaviors. Also, all epochs with wide variation between ORD coder ratings contained unusual driving behaviors. Aspects examined by researchers that did not appear to affect misclassifications included the overall subjective quality of the videos, horizontal curvature of the road, and lighting conditions.

Conclusions

The research indicates that while naturalistic driving data involving time history vehicle data (including vehicle sensors) and face video assessment show promise, they are not currently capable of consistently detecting drowsy driving. With naturalistic driving data, the time history vehicle data algorithm could use the time of day in which the trip began and more pronounced drifts in the lane to screen large data for possible moderately drowsy and very drowsy driving episodes, respectively. For the face video assessment algorithm, although the false alarm and miss rate of the current best performing algorithm would likely make it inappropriate as a warning system for drivers, the performance might be sufficient for screening large naturalistic data sets. While the algorithms may be helpful in locating potential drowsy driving episodes within naturalistic data, significant improvements need to be made before they can be used to consistently identify drowsy driving epochs.

References

National Center for Statistics and Analysis. (2017, October). *Drowsy driving 2015* (Crash•Stats Brief Statistical Summary. Report No. DOT HS 812 446). National Highway Traffic Safety Administration.

National Center for Statistics and Analysis. (2019, October). *2018 fatal motor vehicle crashes: Overview.* (Traffic Safety Facts Research Note. Report No. DOT HS 812 826). National Highway Traffic Safety Administration.

Smith, B. M., Dyer, C. R., Chitturi, M. V., & Lee, J. D. (2017). Automatic driver head state estimation in challenging naturalistic driving videos. *Transportation Research Record*, *2663*(1), 48-56.

Tefft, B. C. (2014). *Prevalence of motor vehicle crashes involving drowsy drivers, United States, 2009-2013*. AAA Foundation for Traffic Safety.

Virginia Tech Transportation Institute. (2015). *SHRP2 researcher dictionary for video reduction data: Version 3.4*.

Virginia Tech Transportation Institute. (2020). *InSight data access website: SHRP2 naturalistic driving study*. https://insight.shrp2nds.us/

Wierwille, W. W., & Ellsworth, L. A. (1994). Evaluation of driver drowsiness by trained observers. *Accident Analysis and Prevention*, *26(5)*, 571-581.

Wiegand, D. M., McClafferty, J., McDonald, S. E., & Hanowski, R. J. (2009). *Development and evaluation of a naturalistic observer rating of drowsiness protocol*. (Report 09-UF-004). Virginia Tech Transportation Institute. https://vtechworks.lib.vt.edu/handle/10919/7412

Suggested APA Format Citation for This Research Note:

National Highway Traffic Safety Administration. (2020, October). *Quantifying drowsy driving* (Traffic Safety Facts Research Note. Report No. DOT HS 813 003).

U.S. Department of Transportation

National Highway Traffic Safety Administration

This research note and other general information on highway traffic safety may be accessed at: www-nrd.nhtsa.dot.gov/CATS/index.aspx