Cascaded Multimodal Analysis of Alertness Related Features for Drivers Safety Applications

Mohamed Abouelenien
Computer Science and Engineering
University of Michigan
zmohamed@umich.edu

Mihai Burzo
Mechanical Engineering
University of Michigan-Flint
mburzo@umich.edu

Rada Mihalcea
Computer Science and Engineering
University of Michigan
mihalcea@umich.edu

ABSTRACT

Drowsy driving has a strong influence on the road traffic safety. Relying on improvements of sensorial technologies, a multimodal approach can provide features that can be more effective in detecting the level of alertness of the drivers. In this paper, we analyze a multimodal alertness dataset that contains physiological, environmental, and vehicular features provided by Ford to determine the effect of following a multimodal approach compared to relying on single modalities. Moreover, we propose a cascaded system that uses sequential feature selection, time-series feature extraction, and decision fusion to capture discriminative patterns in the data. Our experimental results confirm the effectiveness of our system in improving alertness detection rates and provide guidelines of the specific modalities and approaches that can be used for improved alertness detection.

Categories and Subject Descriptors

I.2 [**Artificial Intelligence**]: Miscellaneous

Keywords

alertness; drowsiness; multimodal; cascaded; driving safety

1. INTRODUCTION

According to a survey conducted by the National Sleep Foundation [1], 60% of US adult drivers said they drove while being fatigued, and as many as 37% admitted to have fallen asleep at the wheel. A total of 4% of the drivers had accidents or near accidents due to either drowsiness or falling asleep. In addition to a significantly increased incidence of accidents, a state of drowsiness also has other negative effects: another survey run by the National Sleep Foundation found that from among those driving in a drowsy state, 42% became stressed, 32% became impatient, and 12% drove faster, with all these conditions having negative implications on the safety on the road. An estimate made by the National Highway Traffic Safety Administration indicated that a low alertness level of a vehicle's driver each year results in 100,000 police-reported crashes, which in turn leads to approximately 1,550 deaths, 71,000 injuries, and $12.5 billion in financial damage. However, statistics reported by the National Highway Traffic Safety Administration [3] showed that there existed a decrease of 17.3% in fatal crashes from 1995 to 2012 with the increase in vehicular technology and safety features. According to these sleep studies, a lack of alertness can negatively influence the safety of the drivers and those around them, and therefore a system which can automatically detect and measure the level of alertness of a driver can have a significant positive impact on road safety.

Drowsy driving is a serious problem that can have fatal consequences not only for the driver, but for the passengers and all other road traffic participants (other motorists, pedestrians, and cyclists). With the increase of work load, daily stress, and sleep deprivation, fatigued driving has become very common. Moreover, with the continuous business extension and overnight shipping, driving for extended amounts of time is prevalent.

Despite the presence of studies on the causes affecting drivers' alertness, there is no mean of measuring or detecting their alertness levels. Moreover, driver's license courses and tests do not provide enough information on the importance of keeping alert while driving. In order to reduce the risks associated with drowsy driving, extensive research needs to be conducted to detect the alertness level of vehicles' drivers.

In this paper, we provide a preliminary analysis of the capability of certain modalities on achieving improved detection of drivers' drowsiness, in preparation for collecting a large multimodal alertness dataset that we plan to gather in the near future. The paper presents the results on a dataset provided by Ford which is composed of physiological, environmental, and vehicular modalities. This dataset was provided for the "Stay Alert! The Ford Challenge" in 2011 and is publicly available.[1]

Our analysis is different from previous work as it proposes

[1]Unfortunately, the dataset does not include information on the individual features, and the Ford organizers of the challenge were not able to provide us with any information in addition to what is already included in the data distribution.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org
PETRA '15, July 01 - 03 2015, Island of Corfu, Greece.
Copyright is held by the owner/author(s). Pub. rights licensed to ACM.
Copyright 2015 ACM 978-1-4503-3452-5/15/07... ...$15.00.

a cascaded system that utilizes sequential feature selection, extracts time-series related features, and applies data transformations to acquire patterns that cannot be obtained from raw data. The system analyzes the drivers' status using decision fusion and can employ any type of classifier. The main purpose of this research is to determine which modalities have higher discrimination capability in measuring the alertness level of the drivers and whether integration of multimodal features can induce further improvement. Furthermore, this research provides us with guidelines on which modalities and data types are essential for automated alertness detection. Moreover, we evaluate the sensitivity and specificity of the system to analyze the performance of each of the alertness and drowsiness classes separately to avoid the negative effects of having increased false alarms or missing drowsiness moments. Missed drowsy states can result in traffic accidents while an increased rate of false alarms of drowsiness can lead the driver to ignore serious threats.

This paper is organized as follows. Section 2 discusses related work. Section 3 describes the Ford dataset used in our experiments. Section 4 analyzes the features and details our approach in improving alertness detection. Our experimental results are discussed in Section 5. Finally, the conclusion and guidelines for future work are provided in 6.

2. RELATED WORK

Multiple approaches were used in order to detect drivers' drowsiness using sensorial, physiological, behavioral, visual, and environmental modalities. Earlier methods for drowsiness detection considered the steering motion of vehicles as an indicator of drowsiness [5, 11]. More recent approaches used face monitoring and tracking to detect the alertness levels of drivers. These methods relied on extracting facial features related to yawning [4] and eye closure [9]. Additionally, approaches that measured physiological signals such as brain waves and heart rates were introduced in order to correlate them to states of drowsiness [14].

Kithil et al. [8] developed an alertness detection system using an overhead capacitive sensor array to track the orientation of the head using records of 13 subjects collected at different times through the day. Mao et al. [10] collected physiological measurements such as heart rate, skin conductance, and respiration rate using multiple physiological sensors to detect three drivers' states, namely, alertness, transitional state, and fatigue.

Wahlstrom et al. [18] developed a gaze detection system to determine the activity of drivers in real-time in order to send a warning in states of drowsiness. Gundgurti et al. [6] extracted geometric features of the mouth in addition to features related to head movements to detect drowsiness effectively. However their method suffered from poor illumination and variations in the skin color. Sigari et al. [15] developed an efficient adaptive fuzzy system that derived features from the eyes and the face regions such as eye closure rates and head orientation to estimate the drivers' fatigue level using video sequences captured in laboratory and real conditions. Rahman et al. [12] proposed a progressive haptic alerting system which detected the eyes state using Local Binary Pattern and warned the drivers using a vibration-like alert. Jo et al. [7] introduced a system to separate between drowsiness and distraction using the drivers' faces direction. The system utilized template matching and appearance-based features to detect and validate the location of the eyes using a dataset collected from 22 subjects. However, these individual approaches faced challenges due to errors resulting from signal disambiguation, loss of facial tracking arising from sudden movements, and the invasiveness of signal extraction using contact-based sensors.

Near infrared (IR) spectrum was used to avoid the decrease in the detection rates of eye closure associated with drivers with darker skin and those wearing eyeglasses [16]. Signals such as Electroencephalography (EEG) and electrocardiogram (ECG) were extracted and analyzed to discriminate between the drivers' drowsiness and alertness states [20]. Vezard et al. [17] proposed a genetic algorithm to detect alertness of individuals using EEG signals recorded using 58 electrodes. Linear discriminant analysis and common spatial pattern were employed to select the electrodes that provide the highest alertness detection rate.

Wang et al. [19] surveyed methods introduced to detect drivers' alertness and divided them into methods related to drivers' state such as eyelid movements and percentage of eye closure, methods related to drivers' performance such as distance between vehicles and lane tracking, and multimodal methods that combined both approaches. Doering et al. [13] recorded driving sessions in a driving simulator with the participation of 60 males using a foggy highway to introduce fatigue. Visual and physiological measurements were extracted using a camera and multiple sensors to detect drowsiness.

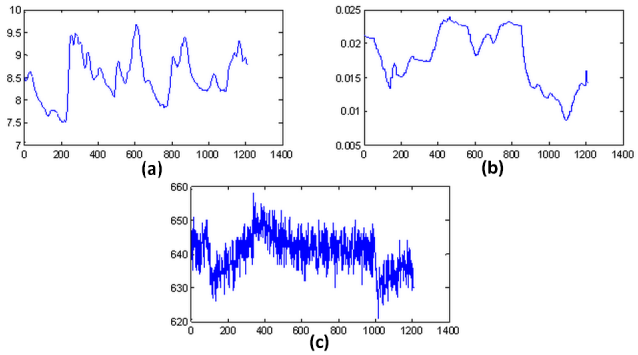

Figure 1: Visualization of multimodal data including (a) a physiological feature, (b) an environmental feature, and (c) a vehicular feature. The figure is provided by Ford on http://www.kaggle.com/c/stayalert/data.

3. DATASET

The publicly available Ford dataset [2], consisted originally of a training set, validation set, and test set of sequential data formed of physiological, environmental, and vehicular modalities. The data was collected from driving sessions on the road and in a driving simulator. Later, the validation and training sets were combined into a single training set. The dataset consisted of recording sessions (trials) for a period of two minutes each, collected from approximately 100 drivers of different ages, ethnic backgrounds, and genders. Sequential measurements were collected every 100 milliseconds during the two minutes trial.

The feature distribution among the three modalities was 8 physiological features, 11 environmental features, and 11 ve-

hicular features. The total number of instances was 604,329 for the training set (formerly a separate training and validation) and 120,840 for the test set, originating from 469 training, 31 validation, and 100 testing trials. Within the same trial, alert and drowsy instances can be found. Visualization of three features of this multimodal dataset can be seen in Figure 1.

Prior to processing the data, we made two interesting observations. First, there were one physiological feature and two vehicular features that had a value of zero among all the instances in the training and test sets. Hence, these three features, namely P8, V7, and V9, were eliminated resulting in a final set of 7 physiological, 11 environmental, and 9 vehicular features. Second, the challenge mentioned Ford's interest in developing a classification model that utilizes fewer physiological features. Our interpretation of this is that the physiological measurements were collected using sensors that were connected to the drivers. Hence, Ford was interested in avoiding the inclusion of additional distractions to the drivers by reducing the number of contact sensors, as well as lowering the costs of the sensors used to detect drowsiness.

4. METHODOLOGY

In order to extract meaningful features to discriminate between states of alertness and drowsiness, we developed a system which employed a cascaded series of feature selection, time-series feature extraction, and classification using decision fusion processes. We used the training and validation sets through each step in this cascade. Finally, our system was evaluated using the test set provided by Ford.

A general diagram of our proposed cascaded system is shown in Figure 2. First feature selection algorithms were applied on the raw data of the physiological, environmental, and vehicular modalities using a sequential search strategy to determine the most discriminative raw features. Second, using the selected features and the fact that they stem from a time-series, we slid windows of predetermined sizes on the data instances to extract time-series related features such as the moving average, maximum, and minimum, the standard deviation, the maximum cross- and auto-correlation, the multilevel approximation coefficient specified by the discrete wavelet transform of the instances in the window, the maximum frequency amplitude of the Fourier transform of the signal, and the power of the frequency domain signal. The idea behind extracting such features was to capture the relationships and dependencies of the data instances and features whether the driver stayed constantly in a single state or switched between alertness and drowsiness.

Third, a model was trained using the final set consisting of the selected raw features combined with the time-series features. Finally, our system was evaluated on the testing set using decision fusion over all window sizes. Each of the previous stages may contain sub-steps to realize its target. Details of each of the aforementioned stages are provided in the following sections.

To achieve our goal in specifying which modality had higher capability of differentiating between alertness and drowsiness, our model was applied on the features of each of the three modalities separately and combined. We additionally specified the percentage contribution of each of the three modalities in the final trained model. Moreover, we calculated the sensitivity and specificity of our model, which were not reported before on this particular dataset. Although we believe detecting drowsy instances to be of higher importance, we decided to select alertness as the positive class and drowsiness as the negative class as declared by Ford in the dataset. The sensitivity metric specified the accuracy of the alertness class (positive class) by dividing the number of true positives by the total number of positive instances. The specificity metric specified the accuracy of the drowsiness class (negative class) by dividing the number of true negatives by the total number of negative instances. False positives can result in missing vital drowsiness periods while increased false negatives can cause frequent false alarms, and hence the driver may ignore serious threats.

5. EXPERIMENTAL RESULTS

5.1 Feature Selection

To specify our feature selection criteria, five different classifiers are trained with the 27 raw features from all three modalities using the training set. The classification models are evaluated using the validation set by measuring the overall classification accuracy. The five classifiers used are decision tree, LibSVM with linear kernel, k-Nearest Neighbor (k=7), Naive Bayes, and Feedforward Backpropagation Neural Network. The decision tree and neural network classifiers are available in MATLAB R2014a. The training and validation data were down-sampled for LibSVM to avoid running out of memory.

Table 1: Percentage accuracy, specificity, and sensitivity achieved by evaluating the validation raw multimodal dataset with all 27 features using five classifiers. The best results are highlighted in bold.

Classifier	Accuracy	Specificity	Sensitivity
Decision Tree	**66.41**	**41.17**	84.57
LibSVM	60.63	6.50	**99.61**
K-NN	60.16	33.827	79.11
Naive Bayes	55.83	32.94	72.30
Neural Network	44.38	11.18	68.28

Based on the results shown in Table 1, a decision tree classifier is selected as the evaluation criterion for the feature selection algorithms. It can also be seen that drowsiness detection has a significantly deteriorated performance compared to the alertness detection, which confirms the importance of further processing the data to avoid the increased false alertness rate.

For feature selection, we employ forward feature selection and backward feature selection methods. The algorithms perform a sequential search strategy to detect the optimal set of features which results in an improved classification model on the validation set. Forward feature selection adds the best single feature at each step based on the evaluation on the validation set. Backward feature selection starts with the total number of features and eliminates the worst performing feature at each step in search for the optimal set.

Both methods are applied on each modality separately and on the whole set of 27 features combined. The selected

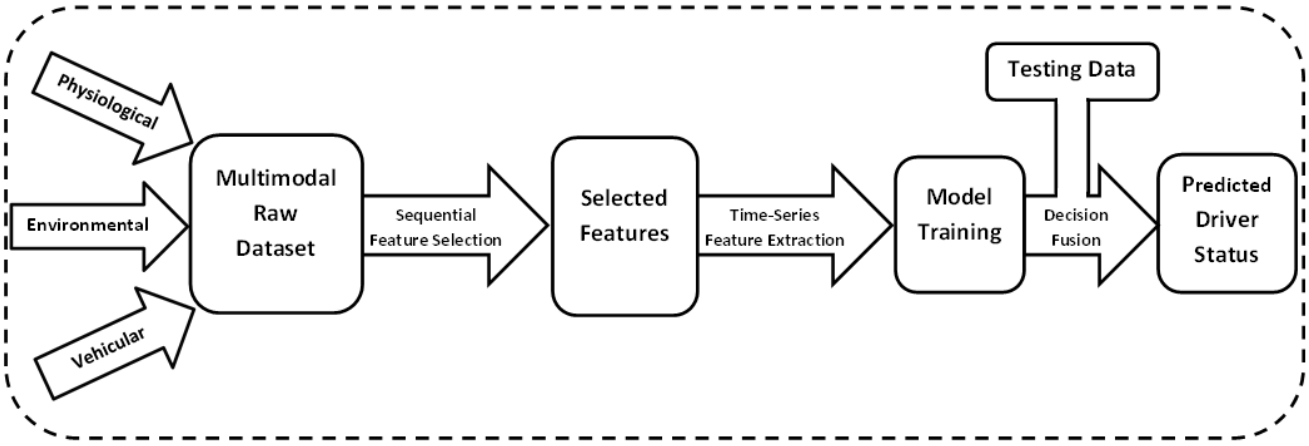

Figure 2: General flow diagram of the major stages of our proposed cascaded system to detect drivers' alertness including sequential feature selection, time-series feature extraction, model training, and final evaluation.

Table 2: Indices of the selected features using forward and backward feature selection as labeled in the dataset provided by Ford before the removal of the zero-valued feature vectors P8, V7, and V9. "P" denotes physiological, "E" denotes environmental, and "V" denotes vehicular modalities.

Dataset	Forward Feature Selection	Backward Feature Selection
Physiological	P5 P6	P5 P7
Environmental	E4 E5 E6 E7 E8 E9 E10 E11	E4 E7 E9
Vehicular	V10 V11	V11
Multimodal	P1 P2 P6 E3 E4 E5 E6 E8 E10 E11 V5 V10 V11	P5 P7 E5 E6 E7 E8 E9 E10 V11

features using both feature selection algorithms are shown in Table 2. Note, however, that the indices provided in the table are the original values as provided by Ford before removing the zero-valued feature vectors P8, V7, and V9 as mentioned earlier.

Table 3 lists the overall accuracy achieved using both feature selection methods applied on individual and combined modalities. "Phys+Env+Veh" presents the fusion of the selected features from each of the physiological, environmental, and vehicular modalities, which were selected separately. "Multimodal" denotes the features selected from all 27 features combined. It can be seen that using backward feature selection on the "Multimodal" features achieves the best overall accuracy outperforming forward feature selection and feature selection on individual modalities. Clearly applying the feature selection algorithms on individual modalities and combining the selected features does not result in detecting the optimal feature set. Hence, the set of features selected using backward feature selection is picked for the feature extraction step for the multimodal dataset. To evaluate individual modalities using our cascaded system, the best set of features for each modality determined in Table 2 and Table 3 are selected, which are P5 and P7 for the physiological, E4, E7, and E9 for the environmental, and V10 and V11 for the vehicular modalities.

It can also be noted that for the multimodal dataset, the majority of the selected features originates from the environmental modality. Two features originate from the physiological and only one is selected from the vehicular modality. The percentage contribution of each of the physiological, environmental, and vehicular modalities in the final set of selected features is 22.22%, 66.67%, and 11.11%, respectively.

Table 3: Overall accuracy percentages achieved using decision tree by training the training set and evaluating the validation set using forward and backward feature selection on individual and combined modalities. "Phys+Env+Veh" presents the fusion of the selected features from each of the individual physiological, environmental, and vehicular modalities. "Multimodal" denotes the features selected from the whole set of 27 features.

Dataset	Forward	Backward
Physiological	66.58	66.58
Environmental	69.27	69.31
Vehicular	68.02	66.93
Phys+Env+Veh	73.67	60.22
Multimodal	74.11	**74.68**

5.2 Time-Series Feature Extraction

In order to extract time-series related features using the set of selected features from the previous step, we slide a window that ends at the current data instance and starts with a lag that is equivalent to the size of the window. The sizes of the windows used in our experiments are 3, 5, 10, 20, and 50. The windows are applied on separate recording sessions or trials. For example, if the current instance within the current trial during the feature extraction process is 120 and the window size is 20, then the time-series features are extracted from instances 101 to 120. Given that the measurements were collected every 100 ms, the windows lags cover a period of 30 ms to 5 seconds.

We create a number of redundant copies of the first instance equivalent to the window size to be able to extract our statistical measurements using lagged instances from the first few instances in a given trial. Using these lagged instances, we extract relationships between different instances as well as features to provide meaningful information on the variations that occur prior to a state of alertness or drowsiness.

Using the group of instances in each window, we extract 10 total time-series related measurements including the moving maximum, minimum, mean, and standard deviation of each of the selected features from the previous stage. These statistical measurements indicate the magnitude of the variations which occurred in each feature during this period of time. By extracting auto- and cross-correlation information from the features, we explore the relations between individual features and lagged copies of themselves as well as the mutual relations between multiple features. Assume the window size is W, the lag range for the cross- and auto-correlation is $2W - 1$. This information can be useful for the training process of our model to discriminate between the relations that lead to an alert or drowsy state.

As we treat a time-series data, extracting frequency-based features can reveal discriminative information which is not directly detected using the time domain signal. In particular, we employ discrete wavelet transform using Haar wavelet and Fourier transform for this purpose. Discrete wavelet transform aims at discovering patterns in the time-frequency domain by decomposing the signal into sub-bands. The resulting coefficients have proven successful in discovering patterns and improving classification results in many applications. For our experiments, we extract the approximation coefficient using the lowest decomposition level. Fourier transform transfers the time-series signal into its frequency components. We extract the maximum frequency amplitude of the signal in a given window in addition to the average signal power distributed over the frequency components. The average power is calculated as the summation of the absolute squares of the signal divided by the length of the window. The first component of the Fourier transformed signal is eliminated as it is the constant component summing up the values in the time-series signal. The feature extraction process is performed on both the training and testing sets. However, it does not take the class labels into consideration. The set of extracted features is appended to the selected features from the previous stage.

5.3 Model Training

The final training set is fed into a classifier for the training process. The same five aforementioned classifiers are used for training. Each of the classifiers is trained five times for the five window sizes, resulting in five different models for each of the five classifiers. The decisions of the five models are combined for each classifier to finally evaluate the testing set.

5.4 Final Evaluation

The testing set is evaluated using each of the trained models of each of the five classifiers: decision tree, LibSVM, K-NN, Naive Bayes, and neural network. The distribution of the training set is 238,882 for the drowsy instances and 328,026 for the alert instances, which creates an imbalance towards the alertness class during training. The distribution of the testing set is 29,914 drowsy instances and 90,926 alert instances, which results in a random guessing baseline performance of 24.76% for the drowsiness class and 75.42% for the alertness class. We evaluate each of the physiological, environmental, and vehicular modalities separately and all the modalities combined. The evaluation metrics include the overall accuracy, specificity, sensitivity, receiver operating characteristic (ROC) curves, and the area under the curve (AUC).

5.4.1 Decision Fusion

The decisions of the models trained using the same classifier with different window sizes are combined using decision fusion. The final prediction of the drivers' state is then determined using the fused model. The decisions are combined using majority voting to determine the final overall accuracy, specificity, and sensitivity. Assume the decision for a given test instance x using each model $f_i(x)$ is given by label $y_i \in \mathcal{Y} = \{1, 2\}$, where 1 is the drowsy class label and 2 is the alert class label. The fused decision is

$$F(x) = \arg\max_{y} \sum_{i=1}^{N} f_i(x) \tag{1}$$

where N is total number of windows; $N = 5$ in our experiments.

To specify the thresholds to create the ROC curves and be able to calculate the AUC, we additionally compute an average score for the fused decisions of different models in the range [1, 2]. The score $S(x)$ for a given test instance x is calculated as

$$S(x) = \frac{1}{N} \sum_{i=1}^{N} f_i(x) \tag{2}$$

5.4.2 Individual Modalities

The individual modalities are evaluated using our cascaded system to determine their separate capability of detecting alertness. In order to determine whether the proposed system improve the performance, we evaluate the testing set using raw data, following feature selection, and following decision fusion. For the feature selection stage, the best set of features for each modality determined in Table 2 and Table 3 are evaluated, which are P5 and P7 for the physiological, E4, E7, and E9 for the environmental, and V10 and V11 for the vehicular modalities. The size of the feature vector for each instance following the feature extraction process for the physiological, environmental, and vehicular modalities is 24, 42, and 24, respectively.

Table 4 evaluates the performance of different modalities using five classifiers by measuring the overall accuracy, speci-

Table 4: The percentage accuracy, specificity, and sensitivity for individual modalities using five classifiers. The best results for each modality-classifier combination are highlighted in bold.

Stage	Physiological			Environmental			Vehicular		
	Accuracy	Specificity	Sensitivity	Accuracy	Specificity	Sensitivity	Accuracy	Specificity	Sensitivity
Decision Tree									
Raw Features	58.42	**30.44**	67.62	**66.33**	**66.34**	**66.33**	61.77	**30.16**	72.17
Feature Selection	59.79	21.79	**72.30**	57.58	65.54	54.96	**68.33**	16.17	**85.49**
Decision Fusion	**60.04**	24.16	71.85	60.82	59.73	61.18	65.08	17.50	80.73
LibSVM									
Raw Features	68.49	**9.27**	87.97	**83.68**	55.37	**93.00**	73.15	**3.78**	95.97
Feature Selection	74.41	0.80	98.62	67.23	44.45	74.73	75.09	0.73	99.55
Decision Fusion	**75.25**	0	**100**	62.93	**61.82**	63.30	**75.25**	0	**100**
K-Nearest Neighbor									
Raw Features	60.39	23.51	72.53	63.22	**66.79**	62.04	64.69	**26.18**	77.36
Feature Selection	62.33	**23.59**	75.08	62.24	59.55	63.12	63.51	18.05	78.47
Decision Fusion	**63.60**	16.75	**79.02**	**64.10**	57.80	**66.17**	**68.49**	12.01	**87.06**
Naive Bayes									
Raw Features	68.38	11.86	86.97	**72.65**	58.90	**77.18**	**42.04**	77.04	**30.53**
Feature Selection	**75.23**	0	**99.99**	62.27	57.54	63.83	39.12	**81.12**	25.30
Decision Fusion	42.87	**39.86**	43.86	60.50	**70.40**	57.24	39.22	80.87	25.52
Neural Network									
Raw Features	75.25	0	100	**78.10**	57.07	**85.02**	71.97	**20.38**	88.94
Feature Selection	75.25	0	100	62.196	**65.48**	61.12	**75.25**	0	**100**
Decision Fusion	75.24	0	99.999	62.78	61.69	63.14	69.78	9.14	89.73

ficity, and sensitivity. It can be noted that the environmental modality provides the best discriminant features to detect alertness. This specific modality also has significantly higher drowsiness detection rates (specified using specificity metric) compared to other modalities which cannot seem to be able to detect drowsiness effectively in most cases.

A problem can be noticed with the usage of the overall accuracy metric with the distribution of the alertness and drowsiness test instances. For example, in several cases for the drowsiness class, the classifier is unable to learn its distribution correctly, and hence classifies all test instances as states of alertness. Yet, the overall accuracy reaches 75% as it is overwhelmed by the large number of alertness testing instances. Therefore, we provided other metrics to fairly assess the performance.

Using the fused decision of our cascade compared to raw features and selected features, the best overall accuracy is achieved 7 out of 15 times for all modality-classifier combinations. In general the physiological and vehicular modalities are close in performance but clearly are not able to discriminate between alertness and drowsiness if used individually. The environmental modality provides reasonable performance, which moves us to the evaluation of the multimodal dataset for comparison.

5.4.3 Integrated Modalities

For the multimodal dataset, the feature selection stage specified nine features: P5, P7, E5, E6, E7, E8, E9, E10, and V11 as the optimal multimodal set. Following the time-series feature extraction stage, each instance has a total size of 153. Each of the 10 extracted measurements provides nine features in addition to the nine raw features, except for the correlation measurements which provide a total of 81 features formed from nine auto-correlation and 72 cross-correlation features.

By comparing the accuracy of the decision fusion stage of each of the five classifiers using the multimodal approach in Table 5 with the accuracy of each individual modality using the corresponding classifier, it is evident that the multimodal approach has a significantly improved performance in all cases except for Naive Bayes classifier. Moreover, the detection rates of the drowsiness class are significantly higher than the baseline without any noticeable drop in the alertness detection rates.

In our experiments, we used predetermined window sizes. In order to analyze the preferred sizes to employ, we compare the accuracy achieved using each window size in Table 6. It is clear that in the majority of the cases, the best performance is achieved by smaller window sizes. This indicates

Table 5: The percentage accuracy, specificity, and sensitivity for the multimodal approach using five classifiers. The top results of each classifier are highlighted in bold.

Stage	Multimodal Dataset		
	Accuracy	Specificity	Sensitivity
Decision Tree			
Raw Features	70.36	60.37	73.64
Feature Selection	72.34	**61.28**	75.99
Decision Fusion	**77.46**	59.09	**83.51**
LibSVM			
Raw Features	87.43	51.23	**99.34**
Feature Selection	**87.49**	**53.07**	98.81
Decision Fusion	83.78	42.371	97.40
K-Nearest Neighbor			
Raw Features	62.77	30.65	73.33
Feature Selection	74.65	**64.79**	77.89
Decision Fusion	**74.86**	61.33	**79.32**
Naive Bayes			
Raw Features	63.65	67.65	62.33
Feature Selection	**73.04**	57.46	**78.17**
Decision Fusion	42.09	**95.26**	24.60
Neural Network			
Raw Features	83.79	**57.73**	92.36
Feature Selection	83.56	52.87	93.66
Decision Fusion	**85.82**	53.29	**96.52**

Table 6: The average accuracy achieved following the feature extraction and prior to the decision fusion stages of the cascaded system using different window sizes.

Dataset	W2	W5	W10	W20	W50
Decision Tree	72.73	73.68	73.75	**74.47**	73.22
LibSVM	**84.06**	74.40	73.54	75.08	75.25
K-NN	**73.86**	73.63	72.85	71.88	70.08
Naive Bayes	**58.75**	53.99	41.15	35.37	35.28
Neural Network	**86.43**	85.79	85.22	84.72	85.03

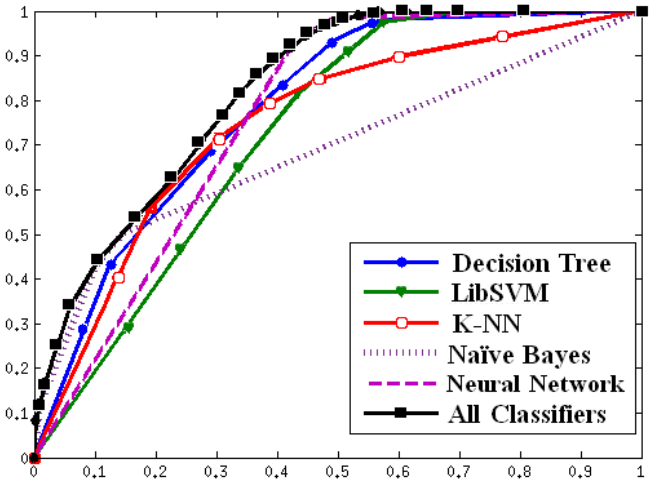

Figure 3: ROC curves following decision fusion of each classifier. "All Classifiers" is formed by further fusing the decisions of all classifiers.

that using a fewer number of lagged instances to extract time-series features can capture useful patterns to separate between states of alertness and drowsiness.

As mentioned earlier, the overall accuracy may not be the best indicator of performance given the test set distribution. Hence, we decided to use the ROC curves and the AUC metric to analyze the performance of our cascaded systems using the scores specified in Equation (2) for each classifier separately. The ROC curves plot the false positive rate (1-specificity) on the x-axis versus the sensitivity on the y-axis. Unlike the accuracy metric, where there can be a high rate of false positives yet high accuracy, the ROC and AUC measure the capability of different classifiers to balance the performance on the alertness and drowsiness classes given their imbalanced distribution. Additionally, we add an optional step in the evaluation process by combining the final decisions made by all types of classifiers using the same equation. This step results in a single overall decision for each instance and is specified as the "All Classifiers" curve in Figure 3. This step is not necessary when the efficiency of the system is of great importance for real-time applications.

Figure 3 shows the ROC curves for each classifier along with the "All Classifiers" curve. The "All Classifiers" and the decision tree curves have the top performance while the Naive Bayes has the lowest performance. To further confirm this observation, Table 7 lists the AUC results of each curve. The best AUC is achieved using the fused classifier reaching 0.8212 and the second best is achieved by decision tree.

6. CONCLUSION

This paper provided an analysis of multimodal features of a dataset provided by Ford in relation to detecting drivers' alertness. The dataset consists of features from physiological, environmental, and vehicular modalities. The main target of this research was to analyze individual and integrated modalities and their capability of discriminating between states of alertness and drowsiness in preparation of collecting a larger multimodal alertness dataset that we plan to gather and make publicly available.

Our experimental results showed that the environmental modality had higher capability of detecting the alertness level of the drivers compared to other individual modalities. However, using environmental features was not sufficient for determining the drivers' alertness state. Integrating selective features from different modalities proved to outperform all single modalities. Hence, we are planning to create

Table 7: AUC results using our cascaded system.

	DT	SVM	KNN	NB	FFNN	All
AUC	0.7830	0.7291	0.7490	0.6873	0.7640	**0.8212**

a multimodal dataset that combines visual, physiological, thermal, environmental, vehicular, and linguistic modalities which, to our knowledge, was never conducted before.

Moreover, we proposed a cascaded system that used sequential feature selection, time-series feature extraction, and decision fusion to effectively detect alertness/drowsiness of the drivers and can be implemented for road safety applications. Evidently, extracting time-series and frequency-related features discovered patterns that were not realized using raw data and resulted in an improved separability between the alertness and drowsiness states. Using windows of lagged instances to extract features, we recommend using windows of smaller sizes as larger windows appeared to blend these patterns especially with rapid switches between alertness and drowsiness states. Furthermore, our proposed system can employ any type of classifier or ensemble of classifiers for a more confident decision.

7. ACKNOWLEDGMENTS

This material is based in part upon work supported by National Science Foundation awards #1344257 and #1355633, and by grant #48503 from the John Templeton Foundation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or the John Templeton Foundation.

8. REFERENCES

[1] National sleep foundation. facts and stats, January 2005. Available at http://drowsydriving.org/about/facts-and-stats/.

[2] Stay alert! the ford challenge, March 2011. Available at http://www.kaggle.com/c/stayalert.

[3] National highway traffic safety administration. national statistics, 2012. Available at http://www-fars.nhtsa.dot.gov/Main/index.aspx.

[4] S. Abtahi, B. Hariri, and S. Shirmohammadi. Driver drowsiness monitoring based on yawning detection. In *2011 IEEE Instrumentation and Measurement Technology Conference*, pages 1–4, May 2011.

[5] L. Barr, H. Howarth, S. Popkin, and R. Carroll. A review and evaluation of emerging driver fatigue detection measures and technologies. Technical report, US Department of Transportation, Federal Motor Carrier Safety Administration, Volpe National Transportation System Center, 2005.

[6] P. Gundgurti, B. Patil, V. Hemadri, and U. Kulkarni. Experimental study on assessment on impact of biometric parameters on drowsiness based on yawning and head movement using support vector machine. *International Journal of Computer Science and Management Research*, 2(5):2576–2580, May 2013.

[7] J. Jo, S. Lee, H. Jung, K. Park, and J. Kim. Vision-based method for detecting driver drowsiness and distraction in driver monitoring system. *Optical Engineering*, 50(12):1–24, December 2011.

[8] P. Kithil, R. Jones, and J. MacCuish. Driver alertness detection research using capacitive sensor array. In *Proceedings of the First International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, January 2001.

[9] S. Kristjansson, J. Stern, T. Brown, and J. Rohrbaugh. Detecting phasic lapses in alertness using pupillometric measures. *Applied Ergonomics*, 40(6):978 – 986, 2009. Psychophysiology in Ergonomics.

[10] Z. Mao, X. Yan, and C. Wu. *Driving Fatigue Identification Method Based on Physiological Signals*, chapter 34, pages 341–352. 2008.

[11] D. Omry and B. Zion. Driver alertness indication system (DAISY). Technical report, Transportation Research Board, November 2006.

[12] A. Rahman, N. Azmi, S. Shirmohammadi, and A. ElSaddik. A novel haptic jacket based alerting scheme in a driver fatigue monitoring system. In *2011 IEEE International Workshop on Haptic Audio Visual Environments and Games (HAVE)*, pages 112–117, October 2011.

[13] M. Rimini-Doering, D. Manstetten, T. Altmueller, U. Ladstaetter, and M. Mahler. Monitoring driver drowsiness and stress in a driving simulator. In *First International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, pages 58–63, August 2001.

[14] A. Sahayadhas, K. Sundaraj, and M. Murugappan. Detecting driver drowsiness based on sensors: A review. *Sensors*, 12(12):16937–16953, 2012.

[15] M. Sigari, M. Fathy, and M. Soryani. A driver face monitoring system for fatigue and distraction detection. *International Journal of Vehicular Technology*, 2013:1–11, 2013.

[16] M. Sigari, M. Pourshahabi, M. Soryani, and M. Fathy. A review on driver face monitoring systems for fatigue and distraction detection. *International Journal of Advanced Science and Technology*, 64(7):73–100, 2014.

[17] L. Vezard, M. Chavent, P. Legrand, F. Faita-Ainseba, and L. Trujillo. Detecting mental states of alertness with genetic algorithm variable selection. In *2013 IEEE Congress on Evolutionary Computation (CEC)*, pages 1247–1254, June 2013.

[18] E. Wahlstrom, O. Masoud, and N. Papanikolopoulos. Vision-based methods for driver monitoring. In *Proceedings of the 2003 IEEE Intelligent Transportation Systems*, volume 2, pages 903–908, October 2003.

[19] Q. Wang, J. Yang, M. Ren, and Y. Zheng. Driver fatigue detection: A survey. In *The Sixth World Congress on Intelligent Control and Automation, WCICA 2006*, volume 2, pages 8587–8591, 2006.

[20] S. Xu, X. Zhao, X. Zhang, and J. Rong. A study of the identification method of driving fatigue based on physiological signals. In *11th International Conference of Chinese Transportation Professionals, ICCTP 2011*, pages 2296–2307, 2011.