Forecasting Markers of Habitual Driving Behaviors Associated With Crash Risk

George Panagopoulos and Ioannis Pavlidis, *Senior Member, IEEE*

***Abstract*— Both distracted and aggressive driving are habitual in nature, constituting an insurance risk, which has been difficult to quantify. Here, in this paper, we propose a method that produces short term predictions for these two dangerous driving behaviors. The method feeds an Extreme Gradient Boosting (XGB) algorithm with the most informative features of a set of physiological and vehicular variables. The XGB algorithm operates on a learning window covering the last 30 seconds to make fast track predictions (FT) for the next 10 seconds. For aggressive driving, FT predictions are final, while for distracted driving, FT predictions are weighted over one minute, to form a meta-prediction. This more deliberative process for predicting distractions fits their intermittent manifestation. The method has been tested on SIM 1, a publicly available dataset from a distracted driving experiment. In this dataset, the drivers ($n = 59$) are labeled as distracted based on the presence of mental activity or physical interactions antagonistic to the driving task; their driving style is defined by steering and acceleration, and is classified as aggressive or normal. The method attains classification performance that exceeds 87%. Alerting drivers when distractions and aggressiveness have taken hold on them can provide sobering awareness, given that people drift into these states subconsciously. The behavioral modification effects of such awareness mechanisms are rooted in Cognitive Behavioral Theory. The proposed method can also be used in future vehicles with advanced automation, weighing in the computer's decision to wrest vehicular control from an unrepentant driver.**

***Index Terms*— Affective computing, distracted driving, aggressive driving, machine learning, extreme gradient boosting, thermal imaging.**

I. Introduction

DRIVING is one of the most common and most dangerous daily human activities. Nearly 1.25 million people die and 20-50 million are injured in vehicle crashes each year worldwide [1]. Several measures to address this problem are implemented from governments and auto manufacturers. The former improve the transportation network, impose stricter penalties, and add further regulations, while the latter equip vehicles with increasingly sophisticated driver assistance systems. Such systems are typically based on cameras and radars to monitor the vehicle surroundings and intervene when an incident happens. Some examples include alerting the driver when departing from lanes, activating the vehicle's breaks on an imminent collision, or warning the driver when the vehicle ahead or behind is too close. These systems have generally made positive contributions to transportation safety. By design, however, they are activated when a serious error has already occurred or is about to occur. Since they solely focus on sensing the position of the vehicle with respect to its surroundings, their ability to foresee drivers' error-prone behaviors, which account approximately for 90% of crashes [2], is limited. Hence, the character of existing driver assistance systems is more remedial than preventive.

This puts beyond the reach of driver assistance systems distracted and aggressive driving, two widespread behaviors of habitual nature, carrying insurance risk that is difficult to document, quantify, and fix. Indeed, there is no good method to sense these two behaviors, and even if there were one, current driver assistance systems cannot take control of the vehicle; thus, remedial actions would solely depend on driver's self-correction. As automation levels advance [3], the role of driving assistance systems is expected to expand from their current 'firefighting' mission to more synergistic actions between the driver and the machine. For example, in vehicles with conditional (Level 3) or high (Level 4) automation, the machine may request or wrest control of the vehicle, if it determines the driver's risky behavior persists despite repeated alerts. For the moment, with vehicles on the road featuring only partial automation (Level 2), a method detecting dangerous behaviors can merely provide sobering warnings to the driver. This warning function should not be underestimated, however, because Cognitive Behavioral Theory suggests it can lead to behavioral modification [4].

Here we present a method that determines when distracted or aggressive driving take hold, issuing alerts. The aim is to improve behaviors by raising driver awareness in Level 2 vehicles. The same method could support preventive actions in Level 3 and Level 4 vehicles. There is a real need for such a development - in 2016 in the United States alone, distracted driving accounted for 3450 human lives [5] while aggressive driving, in the form of oversteering, took 1967 lives [6].

To test the goodness of our method we use SIM 1, an open dataset[1] from a well-known experiment [7] meant to investigate the role of sympathetic arousal in various types of driving distractions [8]. The periods during which the drivers were distracted, are annotated in the dataset at the one second resolution level. The SIM 1 dataset has no annotation for aggressive driving. Hence, we had to develop an annotation scheme based on the definition of aggressive driving, as a combination of excessive steering and acceleration.

Manuscript received May 9, 2018; revised November 13, 2018 and February 4, 2019; accepted March 31, 2019. Date of publication April 23, 2019; date of current version February 3, 2020. This work was supported in part by the Eckhard-Pfeiffer Distinguished Professorship Endowment, and in part by a grant from the Texas A&M Transportation Institute. The Associate Editor for this paper was X. Ma. *(Corresponding author: Ioannis Pavlidis.)*

The authors are with the Computational Physiology Lab, Department of Computer Science, University of Houston, Houston, TX 77204 USA (e-mail: georgepanagopoulos5@yahoo.gr; ipavlidis@uh.edu).

Digital Object Identifier 10.1109/TITS.2019.2910157

[1] https://osf.io/c42cn/

The dataset includes recordings of four different physiological signals for the drivers: Perinasal perspiration, heart rate, breathing rate, and electrodermal activity in the palm. These signals are standard indicators of sympathetic arousal, enabling the detection of overarousal bouts related to cognitive, emotional, or physical overloading of drivers; such overloading increases the risk of accidents [9]. The dataset features also recordings of driving signals from the vehicle's computer; these include acceleration and steering angle signals, serving as indicators of driving style.

We segment the physiological and driving recordings into 10 second windows, within which we perform feature engineering. Subsequently, we use feature selection techniques to arrive at an effective representation of all signals, which feeds a machine learning algorithm - Extreme Gradient Boosting. Both the distraction and aggressiveness models use 30 seconds of recordings to predict whether there will be 'misbehavior' the next 10 seconds. Prediction here is a classification function - the machine learning algorithm gives its classification verdict for the immediate future, having seen the pattern in the recent past. This approach is fitting for persistent behaviors of habitual nature, such as distracted and aggressive driving. Positive classification means that these behaviors have taken hold on the driver and are likely to continue for some time - thus, it is time for an alert.

We call this 30/10 second classification operation, fast track (FT) prediction scale. For aggressiveness, it is the only scale we use. For distractions, however, we treat FT predictions as intermediate results, feeding to a more deliberative scale that averages them over the course of one minute before it issues a meta prediction. This multiscale approach addresses the intermittent nature of distractions, which at the FT level can lead to annoying on/off alert switching. The combination of single scale for aggressiveness and multiscale for distractions yield Area Under the Curve (AUC) in excess of 87%.

The remainder of the paper is organized as follows: In section II we provide a literature review related to this research. In section III we present the methods, which include the dataset, the feature engineering, and the machine learning algorithm. Section IV describes the tests and the results. We conclude in section V with discussion and future work.

II. Related Work

Our work is connected with two different literature sets in driving safety: The first set includes affective monitoring of drivers, while the second set includes classification of instantaneous driving distractions. The goal of affective monitoring during driving is to detect psychophysiological states that lead to dangerous driving behaviors. Most affective driving studies focus on stress and fatigue. Consequently, the definition of stress is a crucial question in these studies. An early research effort categorized driving stress into three classes: no driving, city driving, and highway driving [10]; for validation, the researchers used questionnaires. Along the same lines, another research effort defined stress based on the surroundings [11], with levels corresponding to city, highway, or campus driving. In [12] the subjects drove on a simulated circular driveway for neutral stress, a snowy mountain track for elevated stress, and competed with four artificial intelligence contestants, programmed to induce anger. Stress, however, may be caused by factors other than traffic conditions or the surroundings - a conceptual limitation of early studies.

A more nuanced stress categorization method was used in [13], where stress levels were inferred from driving events - for example, overtakes were associated with elevated stress. In later work from the same group, stress annotations were performed by psychologists, who examined the driving sessions' videos [14]. This approach is feasible in small datasets, featuring few subjects and short driving sessions. It becomes impractical at the scale of the SIM 1 dataset used in the current research. Moreover, even experienced psychologists occasionally fail in assessing people's psychological states by merely looking at their facial videos. To address this issue, researchers resorted to self-annotations from subjects, who were asked to indicate the times they felt frustrated in [15] or tired in [16]. Self-reporting methods, however, have their own set of problems with self-bias being the biggest problem.

In the dataset we use, distraction annotations stem from precise application of stressful stimuli during certain phases of the driving, while controlling or accounting for other factors. These annotations can be considered more objective, because they are free of self-bias or oversimplifying assumptions based on the surroundings. In addition, our dataset includes different types of stimuli that induce either esoteric or physical distractions, thus providing a comprehensive set of scenarios.

Another crucial aspect of driving studies is the data analytic methods. Most driving studies include a limited number of subjects [11], [16], extracting multiple samples from the recordings of each subject; for example, in [14] only 10 subjects are used, but a 10 second sliding window over 40 minute sessions boosts the sampling power. The number of subjects is of vital importance to the model's accuracy, because of the variability found in human beings. Our sample size, with $n = 59$ subjects, exceeds the typical sizes featured in the literature.

Interestingly, while the signals used in driving studies tend to be similar, the preprocessing and feature extraction methods vary widely. In [10], the features used included statistical metrics, power spectral density at different frequencies for respiration and heart rate, as well as peak detection for skin conductance. In [17], researchers used the mean normalized heart rate, the mean absolute first order difference of electrodermal activity, and the mean amplitude of respiration, skin conductance, and facial electromyography. Studies with fewer signals typically focus either on spectral features [16], or peak detection [11]. The frequencies employed in spectral feature engineering differ among studies, and some are not clearly justified. Overall, there is no consensus as to which of these features are the most useful. To address this problem in our study, we start with an extensive set of physiological features that represents the union of the literature. Then, we perform a feature evaluation process to determine the most informative among them. This reduced feature set serves as input to our model.

The machine learning algorithms used in driving studies also play a vital role in modeling effectiveness. With the exception of [10] that used a Fisher's linear discriminant and [11] that experimented with neural networks, the rest of the literature is dominated by Bayesian networks. Bayesian networks can be static [15] or dynamic [17]. Sometimes they are complemented with a first classifier that produces discrete variables from the input, such as a decision tree [17]. Usually, the central variable is binary, representing whether the driver is stressed or not. The rest of the variables can be continuous, such as the measurements of the sensors [15], or nominal, such as types of road events [14]. Bayesian Networks are sound technical choices, because they provide a thorough and explainable probabilistic framework for the model. More importantly, they model the distributions of the inputs, which renders them very robust to overfitting. However, determining the prior probability is an open problem of Bayesian models. In addition, training a Bayesian Network with the wealth of data that our study is based on, can be computationally demanding. This is why in this study we use Extreme Gradient Boosting, a model based on ensembling decision trees, which is not only exceptionally robust to overfitting, but also very fast to train [18].

Overall, our approach differs from traditional distracted driving studies with respect to the type of distractions it targets, the duration of these distractions, and the measurement and analytic methods it uses. Typically, only physical distractions of short duration are considered in driving studies, which are monitored via computer vision systems [19], [20]. These systems employ one or more cameras to track the driver's head pose, mouth movements, and eye blinking, to infer whether s/he gets drowsy or distracted [21]. Simpler systems with head tracking sensors and Controller Area Network-Bus have also been used to classify short distractions [22], such as turning on the vehicle's TV, changing the radio station, or interacting with the navigation system. Instantaneous physical distractions are also detected in [23] but based solely on vehicle signals. The classification method, however, relies on within-subject models to counter inter-subject variance.

In contrast, our approach tackles both physical and mental distractions that persist for some time, following a within- and between-subject design. Such distractions leave a footprint in the subjects' physiological signals and driving behaviors. Our model tries to make sense of this mixed physiological/behavioral footprint, issuing a short term prediction. The aim is to alert the driver in near real-time about behaviors persistent enough to qualify as predictable bad habits. Such a system could serve as the basis of Cognitive Behavioral Therapy for improving driving behaviors [4] and as a monitoring option in auto insurance policies [24]. Table I summarizes the differences between our approach and the models reported in the literature.

III. Methods

A. Dataset

We used the SIM 1 dataset to validate our method. The volunteer subjects had at least one and a half years of driving experience. To control for age, the subjects ranged between 18 and 27 years of age, and above 60 years of age - the young and old groups, respectively. All in all, SIM 1 has $n = 59$ subjects, fairly balanced with respect to gender and age: 12 young males/18 young females plus 14 old males/15 old females.

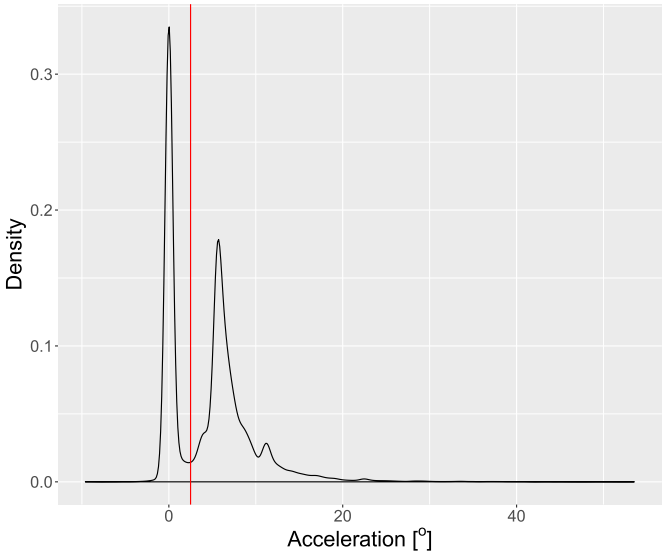

Fig. 1. Probability density function of acceleration for all subjects and drives in the SIM 1 dataset. Acceleration is expressed in °, as the acceleration pedal was connected to a simulated throttle valve that could move from a fully closed (0°) to a fully open (90°) position.

This dataset was produced by an experiment meant to test if esoteric or physical distractions on drivers produce sympathetic overarousal, associated with dangerous changes in driving behaviors [7]. The subjects drove four times the same itinerary in a driving simulator - one time without distractions, and three times with cognitive, emotional, and sensorimotor distractions, respectively. The order of the drives was randomized to ameliorate practice effects. Cognitive distractions were induced with mental arithmetic questions. Emotional distractions were induced with embarrassing questions. Sensorimotor distractions were induced by asking the drivers to text back words that were receiving in their smartphones. The collected data were annotated with the times the distractions started and stopped.

Levels of sympathetic arousal were captured via imaging and wearable sensors recording the following physiological signals:

- Perinasal perspiration through a thermal camera
- Heart rate through a chest sensor
- Breathing rate through a chest sensor
- Electrodermal activity through a palm GSR sensor

At the same time, the vehicle's computer recorded driving signals carrying behavioral information. These signals included instantaneous steering angle and acceleration. The steering angle is a driving parameter closely related to sympathetic arousal, as the arms of the driver handling the steering wheel are the main conduits for funneling 'fight or flight' musculoskeletal responses [8]. In our analysis, we also found acceleration to carry interesting behavioral information, featuring a bimodal distribution (Fig. 1). Naturally, we classify

TABLE I

AN OVERVIEW OF OUR METHOD'S CHARACTERISTICS VERSUS THOSE REPORTED IN THE LITERATURE. 'YES' IN THE CONTROLLED COLUMN INDICATES DATA COLLECTED IN CONTROLLED EXPERIMENTS, WHILE 'NO' SUGGESTS DATA COLLECTED IN FIELD STUDIES

Publication	Distraction	Basis of Annotation	Input	Objective	Response Time [s]	# Subjects	Controlled
[10]	Mental	Environment	Physiological	Detection	300.0	16	**No**
[11]	Mental	Environment	Physiological	Detection	10.0	19	**No**
[12]	Mental	Environment	Physiological	Detection	60.0	20	**No**
[14]	Mental	Expert	Physiological	Detection	10.0	14	**No**
[15]	Mental	Self	Physiological/Behavioral	Detection	0.5	20	**Yes**
[22]	Physical	Movement	Physiological/Behavioral	Detection	0.5	26	**No**
[23]	Physical	Expert	Behavioral	Detection	1.8	20	**Yes**
This work	Mental/Physical	Stimuli	Physiological/Behavioral	Prediction	10.0	59	**Yes**

as 'high acceleration' all the values falling under the right mode, and as 'low acceleration' all the values falling under the left mode in Fig. 1. With respect to steering, we define as 'high steering' all the values that lie beyond one standard deviation in the probability density function of the steering angle variable; values below the one standard deviation threshold are defined as 'low steering'.

We characterize driving to be aggressive in a 10 second window, if both the steering and acceleration values belong to the 'high' classes for at least one second. Our definition of aggressive driving is in the same direction with the definitions used by the insurance industry in the United States. For example, RightTrack® by Liberty Mutual [25] is a telematics rewards program that monitors acceleration signals via an onboard diagnostics (OBD) device; it marks as aggressive driving all incidents with acceleration > 7 miles per hour/second. Our approach, however, differentiates from these insurance-driven programs in two respects: First, our definition of aggressive driving is based on sample statistics, rather than heuristic thresholds, which appears to be the case with RightTrack®. Second, our definition incorporates the latest psychophysiological results in driving research [8], by using hand tremors as an additional indicator of aggressiveness.

Figure 2 depicts the methodological flow of our approach. The remainder of the current section describes in detail the individual elements of this approach.

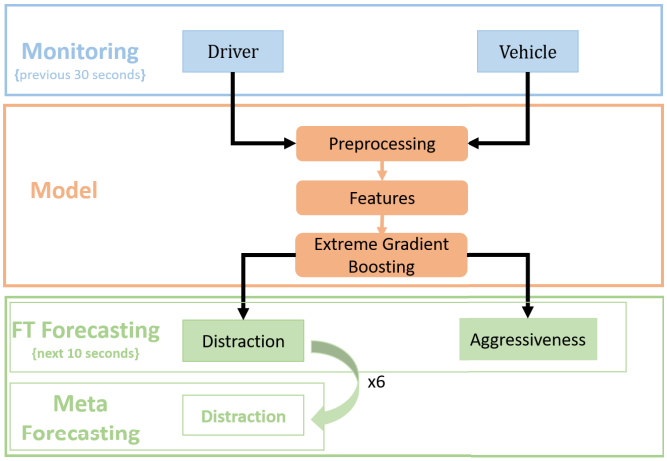

Fig. 2. Overview of methodological flow. FT forecasting stands for fast track forecasting, where data from the last 30 seconds are used to make predictions for the next 10 seconds. Meta forecasting uses six FT forecasting cycles to issue a final prediction; it applies only to distracted driving.

B. Preprocessing

A substantial problem in datasets of such experiments is spurious samples caused by sensor corruption or displacement, as well as human errors. The SIM 1 dataset comes with a detailed quality control report [7], where the experimenters have indicated untrustworthy recordings based on experimental reports, known hardware limitations, and knowledge of physiological boundaries.[2] We utilize this report to filter away corrupted signals. Furthermore, the signals are resampled at 1 Hz, because this is the lowest frequency found in the dataset, coming from the chest sensors measuring heart rate and breathing rate. Finally, we smooth the signals using a simple moving average filter of 5 samples to alleviate artificial variations caused by exogenous factors.

C. Feature Extraction

We perform feature extraction in a time window of 10 seconds. For the driving variables of acceleration and steering, feature extraction is straightforward - we compute their means in a given 10 second window. For the physiological variables, feature extraction is more involved. Different types of physiological responses have different time constants. Cholinergic signals (i.e., perspiration) tend to be highly sensitive, manifesting the onset of stress within 2-5 seconds [26], while adrenergic signals (particularly breathing rate) are slower, needing as much as 10 seconds to manifest arousal [27]. Hence, we choose the upper limit of sensitivity in sympathetic changes (i.e., 10 seconds), as the time window to conduct feature engineering. We extract five different types of features from all physiological signals in a given 10 second window: Statistical, correlative, temporal, structural, and spectral.

1) Statistical Features: These features include the mean, median, standard deviation, sum of squares, and slope of regression. They are standard descriptive statistics capturing the nature of the value distributions in the signals.

2) Correlative Features: Correlations between physiological variables may hide interesting patterns. We extract the upper triangle of the Pearson correlation matrix [28] and the singular values of the covariance matrix.

3) Temporal Features: We assume that signals are stationary within the 10 second observational window. Accordingly, in each time window we extract the following features:

- Auto-correlation parameters [29]
- Coefficients of Autoregressive Integrated Moving Average (ARIMA) [30]
- Sum of squares of ARIMA residuals

[2]https://osf.io/7nwmk/

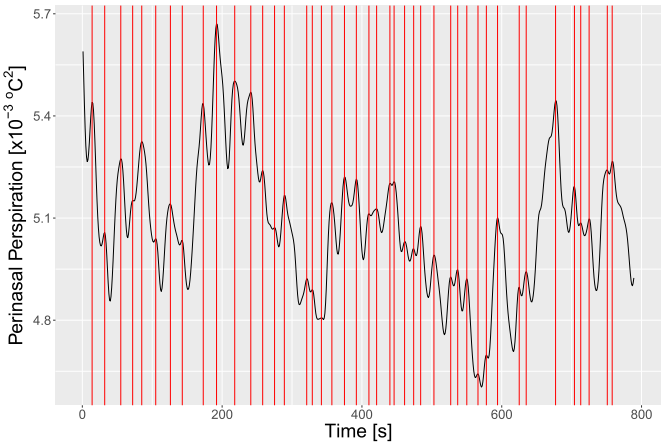

Fig. 3. Example of peak detection in a perinasal perspiration signal (subject S2, driving session with cognitive distractions).

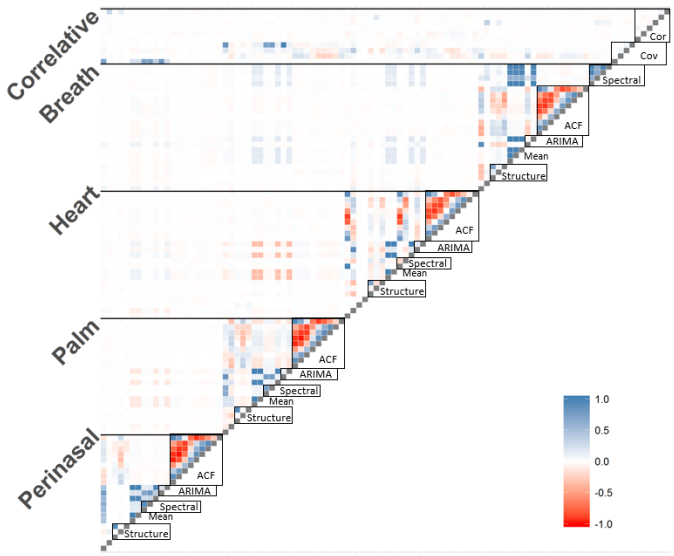

Fig. 4. Heatmap of the feature correlation matrix. Groups of features from the same sensors are separated with horizontal lines and different types of features are highlighted with rectangles. Cor and Cov indicate correlation and covariance of the signal matrix. ACF indicates autocorrelation. Structure indicates features derived from peak detection.

4) Structural Features: These features relate peaks and arousal onsets in cholinergic signals. We use a peak detection algorithm based on zero crossings of the first derivative. The minimum distance between two consecutive peaks is set to 3 seconds, consistent with neurophysiological limitations [26]. The onset of a particular peak is defined as the minimum value between the previous peak and the one currently examined. Figure 3 shows the output of the peak detection for the perinasal perspiration signal of subject S2 in the drive with cognitive distractions. Based on the results of the peak detection algorithm, we extract for each time window the following features:

- Number of peaks
- Mean intensity difference between a peak and its onset
- Mean time between an onset and its peak
- Mean time between consecutive peaks

5) Spectral Features: These are frequency domain features that differ, depending on the type of the physiological signal.

Perinasal Perspiration Perinasal perspiration is a reliable indicator of sympathetic arousal [8]. Like all other signals of cholinergic nature, it can be decomposed into two different components, the tonic and the phasic [31]. Therefore, in each time window we extract the maximum value of the spectrogram in $[0-0.2]$ Hz, the maximum value of the spectrogram in $(0.2, \infty]$ Hz, and the respective frequencies of these two maxima.

Palm Electrodermal Activity. Electrodermal activity (EDA), much like perinasal perspiration, is a cholinergic signal [31]. Hence, we extract the same spectral features we extracted for the perinasal perspiration signal.

Heart Rate. We use Lomb periodogram, because it is robust to missing samples, to extract frequency components that capture the influence of the sympathetic and parasympathetic system on the heart. These features include power spectral density in (0.003 - 0.04], (0.04 - 0.15], (0.15 - 0.5] Hz, as well as their ratio, and the total power spectral density.

Breathing Rate. We use Welch's average periodogram with Hanning window in the frequencies: (0, 0.1], (0.1, 0.2], (0.2, 0.3], and (0.3, 0.4] Hz.

D. Feature Selection

The total number of extracted features amounts to 111. To lessen the possibility of overfitting we reduce the dimensionality of the feature set. Traditional dimensionality reduction techniques, such as principal component analysis, although effective, are not desirable in this case, because we want to maintain some control in the process, keeping the physical meaning of our model in perspective. Consequently, we resort to feature selection techniques.

First, we remove features with constant values. These features cannot contribute to class discrimination, as they do not change at all. This filtering eliminates all first autocorrelation parameters, as well as the maximum frequency components of the perinasal and palm EDA signals. Afterwards, we deal with the unavailable values (N/A) left behind by the feature extraction process. We remove features that consist of more than 10% N/A values. This multi-step cleaning process results into a dataset with 98 features and 43 subjects, forming a total of 9956 distinct samples.

Subsequently, we apply min-max normalization to reduce differences between various types of features with aberrant scales. Next, we remove features that are highly correlated with other features, that is, they exhibit over 0.8 positive or negative Pearson correlation. Figure 4 shows the feature correlation matrix. The matrix follows a block structure across the opposite diagonal. Using the horizontal lines that highlight the features of a specific sensor, we can conclude that features from the same sensors can be highly correlated with each other, but the correlations between features from different sensors are scant.

Focusing on each modality, we see a triangle with high negative and positive correlations. These correspond to the autocorrelation parameters, each one being positively correlated with the immediate previous ones, and negatively with the ones that are further back These correlative relationships cause the removal of a significant number of features in each sensor. The correlation of spectral density features depends on the sensor modality. For example, the two maximum spectrum features of palm EDA are not correlated with any other feature, while the respective perinasal ones are found to be correlated with each other. For all modalities, the ARIMA residuals are positively correlated with the standard deviation, while the ARIMA parameters are positively correlated with the mean. This makes sense as the residuals relate to variance, and the ARIMA parameters rely on mean estimates. Means are also positively correlated with energy and medians, indicating that the signal distributions are not skewed. Mean breathing correlates with several spectral features of breathing. Out of all these correlated features, we keep just the means as they are more robust and easier to compute. The most dominant and weakest eigenvalues of covariance, are positively correlated with the mean of palm EDA and the mean of perinasal, respectively. This suggests that palm EDA is the strongest contributor in signal variance, which, given that it is our most noisy signal, makes sense. To verify that strong correlative indications do not stem from pseudo-correlations, we also plotted each highly correlated couple for visual inspection. 48 features were left standing after this process.

Finally, we employ feature selection methods from the machine learning literature to trim the feature set down to its most informative components with respect to predicting distraction and aggressiveness. Towards this, we use two methods. The first method is based on the feature importance index computed by Extreme Gradient Boosting machines [18]. The task is to predict distraction using all 48 features and keep the 10 most informative ones. In the second method, we use simple linear models where each feature acts as a sole predictor of distraction. We use 10-fold cross-validation to identify the 10 most successful predictors. We also use these two methods to predict aggressiveness. We take the intersection of the resulting four feature vectors to form an optimal set of 17 physiological features. The final feature vector has these 17 physiological features plus the two mean driving features (Table II). Interestingly, there is 80% overlap between the best predictors for distraction and aggressiveness, indicating a common physiological and behavioral core for these two dangerous habits.

TABLE II

FINAL SET OF FEATURES EMPLOYED IN THE MODELS

Mean and Standard Deviation of Perinasal Perspiration
Mean and Standard Deviation of Heart Rate
Mean and Standard Deviation of Palm EDA
Mean and Standard Deviation of Breathing Rate
Maximum Spectral Density of Perinasal Perspiration
First and Second Maximum Spectral Density of Palm EDA
Very High Frequency Spectral Density of Breathing Rate
Second and Third Eigenvalue of Covariance Matrix
Ratio of Low to High Frequency Spectral Density of Heart Rate
Rate of Change of Palm EDA
Correlation between Perinasal Perspiration and Breathing Rate
Mean Acceleration and Steering

TABLE III

COMPARISON OF XGB VERSUS NAIVE BAYES, GENERALIZED LINEAR MODEL (GLM), AND SUPPORT VECTOR MACHINES (SVM) IN THE SIM 1 DATASET. THE AREA UNDER THE CURVE (AUC) AND ACCURACY (ACC) PERFORMANCE SCORES DEMONSTRATE THE SUPERIORITY OF XGB IN PREDICTING DISTRACTIONS AND AGGRESSIVENESS AT THE FT LEVEL

		XGB	Naive Bayes	GLM	SVM
DISTRACTIONS	AUC	**84.26**	56.6	67.2	73.0
	ACC	**78.36**	61.9	69.1	70.5
AGGRESSIVENESS	AUC	**87.13**	50.9	76.0	81.3
	ACC	**89.24**	53.7	87.0	88.0

E. Extreme Gradient Boosting

We chose Extreme Gradient Boosting (XGB) because of its robustness in overfitting, its interpretability, and its computational efficiency [18]. Furthermore, XGB can handle relatively successfully unbalanced class problems such as ours; distractions have a 1:2 positive-negative class ratio and aggressiveness has 1.5:8.5. Indeed, XGB surpassed in performance well-known machine learning algorithms in a comparative experiment we ran against the SIM 1 dataset (Table III).

The XGB model is an ensemble of decision trees, meaning that during training multiple trees are constructed sequentially in a stepwise manner, each taking into account the weaknesses of the previous one. When testing a new sample, each tree gives a probability score for each class and a weighted combination of them gives the final estimate. A gradient descent methodology is adapted to optimize the structure of the trees in each step. The vector of predictions of the model in step t is defined as:

$$\hat{y}^t = \sum_{k=1}^{t} f_k(x) = \sum_{k=1}^{t-1} f_k(x) + f_t x = \hat{y_i}^{t-1} + f_t(x_i) \quad (1)$$

where $f_k(x)$ represents the function corresponding to the tree developed at step k, with T leaves and $w \in R^T$ being the scores assigned to the samples in those leaves. The optimization aim is to reduce the binary loss from $\hat{y}^t$ to the real value of y:

$$l(y, \hat{y}^t) = \sum_{i=1}^{n} (y_i ln(1 + e^{-y_i^t}) + (1 - y_i) ln(1 + e^{y_i^t}) \quad (2)$$

and the regularization term, which serves as a measure for the complexity of the model, is defined as:

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \quad (3)$$

where γ and λ are hyper parameters, and w_j is the score of the j-th leaf. Adding Eq. (2) and Eq. (3) gives the objective function, which measures how good the structure of the tree is. By solving this function the optimum score of j-th leaf is:

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \quad (4)$$

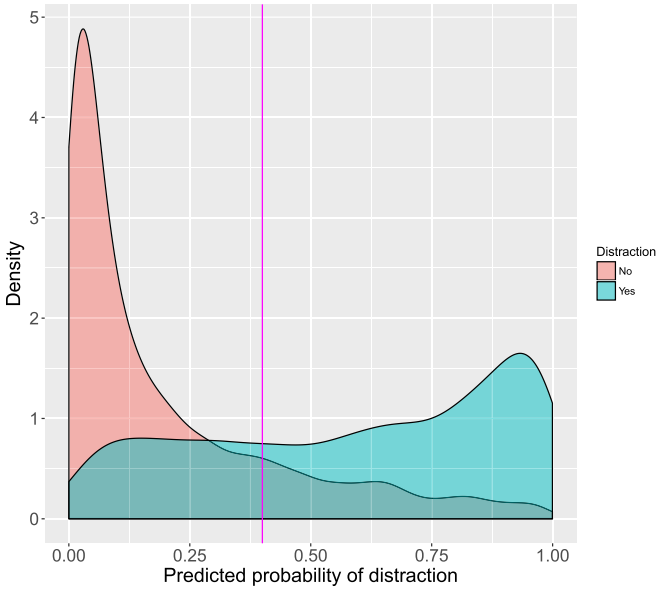

Fig. 5. The density of the model's probabilistic distraction predictions, color-annotated with ground-truth information.

where I_j are the indices of the samples classified to that leaf and g_i, h_i are the first and second derivatives of the loss $l(y, \hat{y}^{t-1})$. To employ this when constructing the trees, a gain index is defined to estimate what is the best feature to choose as the next decision node. This index depends on the leaves created by the candidate decision node (R and L) and the gain on the current node I:

$$gain = -\frac{1}{2}[w_R^* + w_L^* - w_I^*]. \quad (5)$$

F. Multiscale Predictions

The method makes predictions at two time scales - a fast track (FT) prediction and a meta prediction after a period of deliberation. The FT prediction applies to both distracted and aggressive driving. Predictions are issued for the next 10 seconds (x) using 30 seconds ($3x$) from the immediate past. Hence, the sampling process operates on a 40 second time window that slides forward. The lag has been set to 3x, after sensitivity analysis ranging from $1x$ to $10x$.

For aggressive driving, the FT prediction is the only prediction that takes place. For distracted driving, the FT prediction serves as an intermediate layer, feeding to a meta prediction mechanism. This higher level process averages FT predictions for six cycles (i.e., 60 seconds in total), before issuing an ultimate prediction for distracted driving the next 10 seconds. The probabilistic threshold for the meta-predictor to issue a distraction 'alarm' has been set to 0.4. The optimality of this threshold is evident in Fig. 5, which shows the probabilities of distraction given by the model, color-annotated with the ground-truth information.

Texting while driving and smartphone distractions in general are habitual and persistent in nature, characterized by an intermittent pattern that lasts a lot longer than a few seconds. The same applies for esoteric distractions, such as absent-mindedness, especially in the context of tedious commutes. Hence, it makes sense to have a multiscale prediction scheme for the said distractions, because their typical time frame would support it. Taking a more deliberative approach would likely increase the accuracy and reliability of the system, as it is likely to avoid getting caught in local minima created by the intermittent pattern of distractions.

In contrast to distractions, aggressive driving events are characterized by relatively shorter durations and have bursty nature. A key reason is that acceleration is an important determinant of aggressive driving, and usually one cannot keep accelerating/decelerating for very long or very often. Hence, prediction of aggressive driving is best to operate at the FT level only.

TABLE IV

5-FOLD CROSS VALIDATED EVALUATION METRICS AND RESULTS FOR FAST TRACK (FT) PREDICTIONS FOR DISTRACTIONS AND AGGRESSIVENESS (EVERY 10 s), AND META-PREDICTIONS (MP) FOR DISTRACTIONS ONLY (EVERY 1 min): AREA UNDER CURVE (AUC), ACCURACY (ACC), SPECIFICITY (SPEC), SENSITIVITY (SENS), POSITIVE PREDICTIVE VALUE (PPV), NEGATIVE PREDICTIVE VALUE (NPV)

	Distractions (FT)	Distractions (MP)	Aggressiveness (FT)
AUC	84.26	88.74	87.13
ACC	78.36	80.82	89.24
SPEC	87.28	84.87	95.52
SENS	61.47	73.49	48.84
PPV	71.83	72.84	62.88
NPV	81.10	85.29	92.32

IV. RESULTS

To evaluate the effectiveness of the method, we apply 5-fold cross-validation [32] both at the fast track (FT) and meta prediction levels. As part of this process, we keep ∼2000 samples for testing and we use the remaining ∼8000 samples for training. This is repeated for five times, choosing a different subset of ∼2000 samples each time to produce predictions.

Table IV shows that the Area Under the Curve (AUC) is over 84% for both distraction and aggressiveness predictions, across scales. AUC, which is based on the rate of true versus false positives [33], is more reliable than binary accuracy in this application due to the heavy imbalance of classes. Indeed, drivers are non-distracted and drive non-aggressively most of the time. This class imbalance contributes to the disparity between specificity and sensitivity, especially at the FT level. The high specificity and high negative predictive value show that the model is very good at predicting when drivers are not distracted or not aggressive. The FT sensitivity values show that the model underpredicts true positive states. The combination of these qualities is revealing of a model that will alert drivers on rare occasions and will usually be right. This is a requirement for a driving behavioral model to be successful. Drivers would not be keen on using systems that admonish them often and for no good reason. Hence, reliably predicting when not to warn drivers, is crucial.

Figure 6 provides an insight into the model's performance. Specifically, Fig. 6 A, B visualize the model's meta and FT level performance in predicting distractions. Figure 6 C

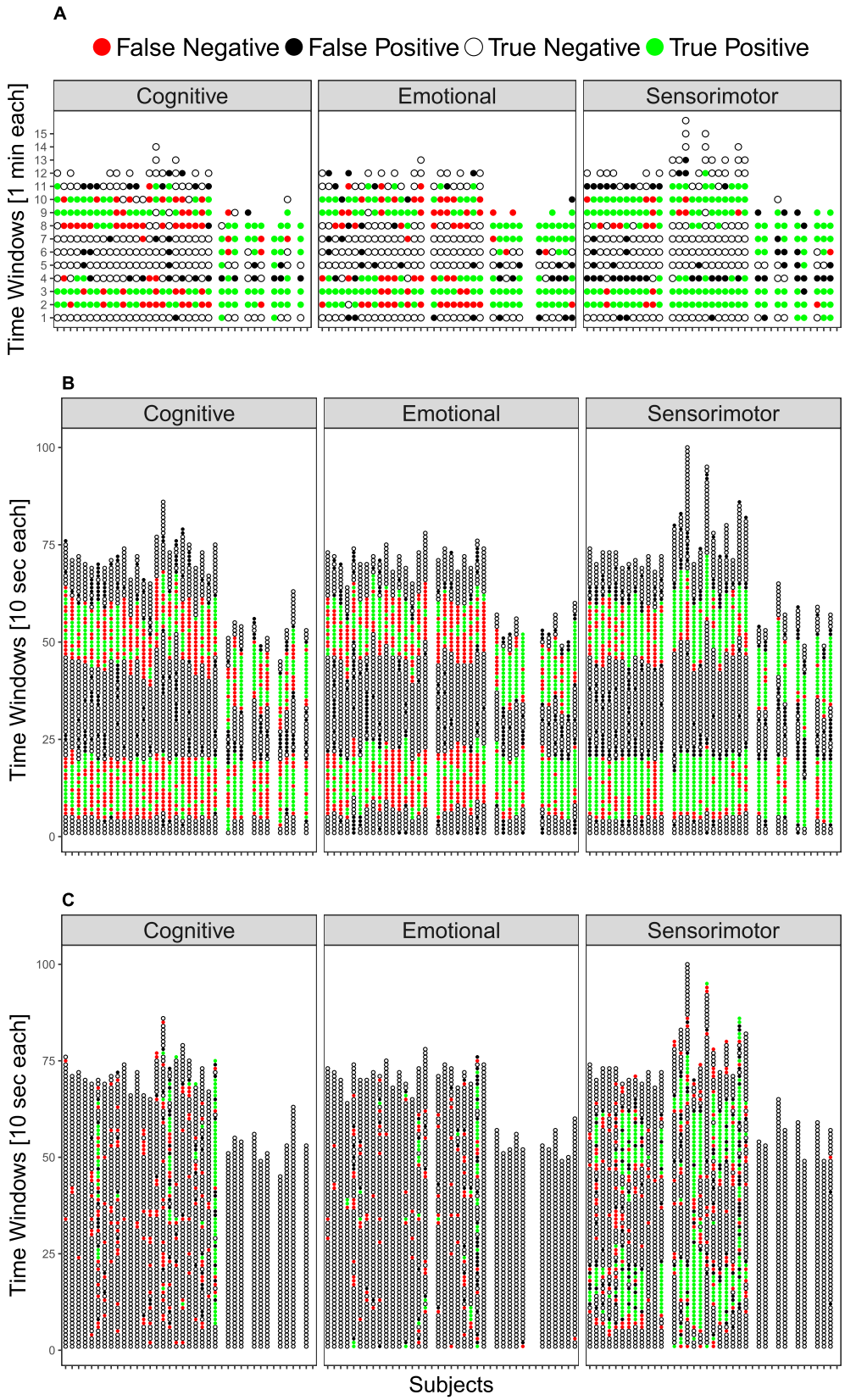

Fig. 6. Visualization of the model's predictive results. **A.** Meta predictions for distractions. **B.** Fast track (FT) predictions for distractions. **C.** Fast track predictions for aggressiveness.

visualizes the model's FT level performance in predicting aggressiveness. Each plot corresponds to a drive afflicted by a specific type of distraction, per the SIM 1 experimental design. Green discs denote true positive predictions, black discs denote false positive predictions, and red discs denote false negative predictions. Hollow discs denote true negatives - clearly, true

negatives constitute the majority of states, and thus the class imbalance discussed earlier. In Fig. 6 A,B, green discs cluster in two phases; before, in between, and after these two phases there is preponderance of hollow discs. This reflects the SIM 1 experimental design [8], suggesting that the model captures well the overall pattern of driving behaviors.

At the FT scale, the false negatives are dispersed within the phases that the subjects were cognitively, emotionally, or physically distracted. This happens because distractions are not 100% on all the time. They rather have an interlacing pattern, where, for example, in a 10 minute period, thoughts are coming in and out of a driver's mind, or the driver texts intermittently. Interestingly, the false positives at the FT scale come mostly after the end of the distracted driving phases, representing after-effects that shortly outlive the stressors, as documented in the literature [8]. Hence, labeling has some limitations here, and in terms of substance, some of the false negatives and false positives are likely not false at all. Furthermore, this conundrum also points to a practical problem: Assume that we had a perfect method to label short non-distracted intervals within a period of overall distracted driving - would it be proper or meaningful to continually switch on and off the alerts?

This analysis prompted us to consider a more deliberative decision time scale for issuing distraction alerts. We implemented this as a meta process, weighing on six FT predictions that operate as a hidden layer. Meta predictions of distractions significantly improved performance across all metrics and particularly with respect to sensitivity.

Some of the patterns identified in the prediction of distractions are also present in the prediction of aggressiveness. These patterns manifest more sparsely in the drives with esoteric distractions (cognitive and emotional), and more intensely in the drive with physical distractions (sensorimotor). This is consistent with previous findings, which documented that the anterior cingulate cortex (ACC) manages subconsciously (and effectively) the driving function when pure esoteric distractions are at play [8]. In the context of the SIM 1 dataset that features straight highways and little traffic, this makes for a smooth ride. ACC, however, fails in its function when physical distractions are introduced [8], resulting in lane deviations for which the driver needs to take corrective action, resulting in acceleration adjustments and significant steering corrections, all of which drive up the aggressiveness factor.

The model draws on two types of variables to perform its machine learning and prediction functions - physiological variables from the imaging and wearable sensors, and driving variables from the vehicle's computer. To quantify the relative usefulness of each type of variable, we ran a comparative experiment feeding the model once with the physiological variables, once with the driving variables, and once with both. Figure 7 shows the results, confirming the valuable role of the physiological variable set, especially in the prediction of distractions, where it outperforms the driving variable set, with the exception of sensitivity. The combined set of variables (physiology + driving) nearly always outperforms each individual set, suggesting a level of complementarity. This outcome supports previous findings regarding the use

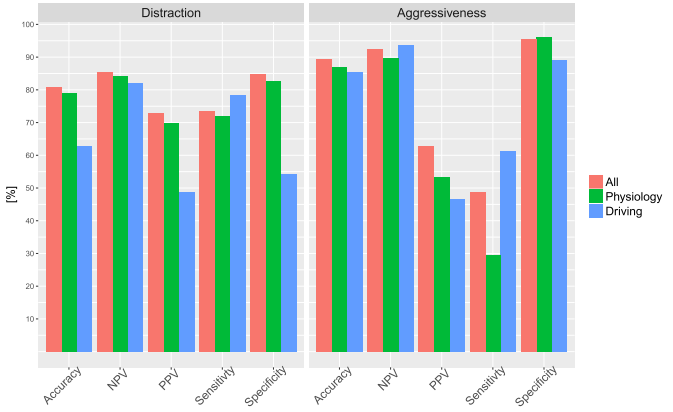

Fig. 7. Relative contribution of the physiological and driving features to the model's performance.

of physiological and performance variables for classifying cognitive workload during driving [34].

In support of open science, we made available an application in R that visualizes predictions vs. ground truth per subject and experimental session.[3] This application also gives quantitative estimates of the model's success per case.

V. Discussion

This research produced a model that alerts drivers when dangerous driving behaviors of habitual nature appear to take hold. The work is motivated by the increasingly grim statistics of crashes due to distractions, as well as the tedious commutes in metropolitan areas that tend to throw drivers into absent-mindedness or aggressive responses. In the near future, the aim is for the system to act either as a behavioral orthotic, or as a quantitative risk assessment tool for insurance companies. Further down the line, when vehicles with Level 3 and Level 4 automation become reality, such a system could factor in the machine's decision to wrest control of the vehicle.

The model uses physiological and driving variables within a machine learning context to issue fast track (FT) predictions for distracted or aggressive driving every 10 seconds. These predictions are based on the evolution of the physiological and driving variables the past 30 seconds. While for aggressiveness FT predictions are the final predictions, for distractions, FT predictions serve as hidden intermediary indicators that are weighed over a minute to form a meta prediction. This more deliberative approach addresses the intermittent patterns of distractions. Such patterns can lead FT predictions to misdiagnose a pause in a period of distracted driving as the end of it, resulting in frequent on/off alert switching. Hence, in classification terms, FT predictions result in lower sensitivity. In practical terms, FT predictions undermine the driver's trust to the system, exhausting her/his good will.

This investigation documented that both physiological and driving variables play a key role in predicting distractions and aggressiveness; their combination nearly always improves classification results. This is partly due to the complementary nature of these two variable sets - the physiological set excels in accuracy and PPV, while the driving set is strong in terms of sensitivity and NPV (Fig. 7).

[3]https://georgepanagopoulos.shinyapps.io/ForecastRoadBehavior/

Regarding the feasibility and practicality of the system, the core portion of it is immediately implementable in current Level 2 vehicles. All the driving parameters the method uses (i.e., steering angle and acceleration) are recorded by the computer of a typical Level 2 vehicle. Regarding the physiological variables, some are recoverable from the drivers' smart watches (i.e., heart rate), which are connected to the vehicle's computer via Bluetooth. Some other more exotic physiological variables, such as perinasal perspiration from thermal imaging, have not been commoditized yet. However, extrapolating the price drop in thermal imaging sensors the last decade, it is almost certain that such measurements will be commoditized by the time Level 3 and Level 4 vehicles are on the streets.

Using physiological variables to detect the affective state of drivers goes back in time [35]. However, combining physiological with driving variables is much less common. Furthermore, in comparison with other prior work [34], the present method has the advantage of being subject independent. It also relies on a sizable, well-abstracted, and validated dataset that includes both esoteric and physical distractions [8]. This is in contradistinction to the typically small, ad-hoc (i.e., one type of distraction), and non-validated datasets upon which other methods were trained and tested [36].

One limitation of the current dataset is with respect to aggressive driving incidents. Such incidents are scant because the SIM 1 study design that produced the data was focused on distracted driving. As a result, the aggressive driving prediction method is trained on highly unbalanced data and the resulting sensitivity is low. To appreciate how the sensitivity of the aggressive driving predictor would scale up in a more appropriate dataset, one has to look at the sensitivity of the distracted driving predictor in the present dataset. Irrespective of this, the specificity of the system is excellent, which means that the method may be missing some aggressive driving events, but whenever it issues aggressive driving notifications, is almost always right. Consequently, the method will appear trustworthy to the driver, which is key to acceptance and behavioral modification.

All in all, the presented distracted and aggressive driving prediction models demonstrated very good performance in state of the art controlled experiments. Future enhancements would benefit from testing on a naturalistic dataset, where other factors, such as weather and traffic conditions, would need to be incorporated into the models. The incorporation of additional sensory measurements from drivers' smartphones also hold promise [37], although relevant validated datasets are for the moment in short supply. Last but not least, future investigations should also focus on sharpening the machine learning aspects of the method. In this respect, long short-term memory neural networks [38] could result in higher accuracy. Multi-task learning is also worth exploring. Several researchers have employed multi-task learning models to overcome cross subject generalization problems [39].

ACKNOWLEDGMENTS

The authors would like to thank Dr. Panagiotis Tsiamyrtzis and Dr. Vangelis Karkaletsis for their help. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the sponsoring entities.

REFERENCES

[1] Association For Safe International Road Travel. (2018) *Road Safety Facts*. [Online]. Available: https://www.asirt.org/safe-travel/road-safety-facts/

[2] The Center for Internet and Society. (2018) *Human Error as a Cause of Vehicle Crashes*. [Online]. Available: http://cyberlaw.stanford.edu/blog/2013/12/human-error-cause-vehicle-crashes

[3] On-Road Automated Driving Committee, "Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles," SAE Int., Pennsylvania, PA, USA, Tech. Rep. J3016, Jan. 2014.

[4] P. C. Kendall and S. D. Hollon, *Cognitive-Behavioral Interventions: Theory, Research, and Procedures*, vol. 21. New York, NY, USA: Academic, 2013.

[5] National Highway Traffic Safety Administration. (2017). *Distracted Driving*. [Online]. Available: https://www.nhtsa.gov/risky-driving/distracted-driving

[6] Insurance Information Institute. (2017). *Facts + Statistics: Aggressive Driving*. [Online]. Available: https://www.iii.org/fact-statistic/facts-statistics-aggressive-driving

[7] S. Taamneh *et al.*, "A multimodal dataset for various forms of distracted driving," *Sci. Data*, vol. 4, Aug. 2017, Art. no. 170110.

[8] I. Pavlidis *et al.*, "Dissecting driver behaviors under cognitive, emotional, sensorimotor, and mixed stressors," *Sci. Rep.*, vol. 6, May 2016, Art. no. 25651.

[9] F. P. McKenna, "The human factor in driving accidents An overview of approaches and problems," *Ergonomics*, vol. 25, no. 10, pp. 867–877, 1982.

[10] J. A. Healey and R. W. Picard, "Detecting stress during real-world driving tasks using physiological sensors," *IEEE Trans. Intell. Transp. Syst.*, vol. 6, no. 2, pp. 156–166, Jun. 2005.

[11] R. R. Singh, S. Conjeti, and R. Banerjee, "A comparative evaluation of neural network classifiers for stress level analysis of automotive drivers using physiological signals," *Biomed. Signal Process. Control*, vol. 8, no. 6, pp. 740–754, 2013.

[12] J. S. K. Ooi, S. A. Ahmad, Y. Z. Chong, S. H. M. Ali, G. Ai, and H. Wagatsuma, "Driver emotion recognition framework based on electrodermal activity measurements during simulated driving conditions," in *Proc. IEEE EMBS Conf. Biomed. Eng. Sci. (IECBES)*, Dec. 2016, pp. 365–369.

[13] G. Rigas, C. D. Katsis, P. Bougia, and D. I. Fotiadis, "A reasoning-based framework for car driver's stress prediction," in *Proc. 16th Medit. Conf. Control Autom.*, 2008, pp. 627–632.

[14] C. D. Katsis, Y. Goletsis, G. Rigas, and D. Fotiadis, "A wearable system for the affective monitoring of car racing drivers during simulated conditions," *Transp. Res. C, Emerg. Technol.*, vol. 19, no. 3, pp. 541–551, 2011.

[15] L. Malta, C. Miyajima, N. Kitaoka, and K. Takeda, "Analysis of real-world driver's frustration," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 1, pp. 109–118, Mar. 2011.

[16] B.-G. Lee and W.-Y. Chung, "Driver alertness monitoring using fusion of facial features and bio-signals," *IEEE Sensors J.*, vol. 12, no. 7, pp. 2416–2422, Jul. 2012.

[17] G. Rigas, Y. Goletsis, and D. I. Fotiadis, "Real-time driver's stress event detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 1, pp. 221–234, Mar. 2012.

[18] T. Chen and T. He, *Xgboost: Extreme Gradient Boosting*, document R package 0.82.1, 2019.

[19] M. Kutila, M. Jokela, G. Markkula, and M. R. Rue, "Driver distraction detection with a camera vision system," in *Proc. IEEE Int. Conf. Image Process.*, vol. 6, Sep./Oct. 2007, pp. VI-201–VI-204.

[20] W. Rongben, G. Lie, T. Bingliang, and J. Lisheng, "Monitoring mouth movement for driver fatigue or distraction with one camera," in *Proc. 7th Int. IEEE Conf. Intell. Transp. Syst.*, Oct. 2004, pp. 314–319.

[21] C.-W. You *et al.*, "CarSafe: A driver safety App that detects dangerous-driving behavior using dual-cameras on smartphones," in *Proc. ACM Conf. Ubiquitous Comput.*, Sep. 2012, pp. 671–672.

[22] M. Wollmer *et al.*, "Online driver distraction detection using long short-term memory," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 2, pp. 574–582, Jun. 2011.

[23] F. Tango and M. Botta, "Real-time detection system of driver distraction using machine learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 2, pp. 894–905, Jun. 2013.
[24] Y. Dong, Z. Hu, K. Uchimura, and N. Murayama, "Driver inattention monitoring system for intelligent vehicles: A review," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 2, pp. 596–614, Jun. 2011.
[25] Liberty Mutual Insurance. *Welcome to RightTrack by Liberty Mutual*. Accessed: Apr. 2, 2019. [Online]. Available: https://www.libertymutual.com/righttrack
[26] P. Tsiamyrtzis, M. Dcosta, D. Shastri, E. Prasad, and I. Pavlidis, "Delineating the operational envelope of mobile and conventional eda sensing on key body locations," in *Proc. CHI Conf. Hum. Factors Comput. Syst.* New York, NY, USA: ACM, 2016, pp. 5665–5674.
[27] I. Pavlidis *et al.*, "Fast by nature-how stress patterns define human experience and performance in dexterous tasks," *Sci. Rep.*, vol. 2, p. 305, Mar. 2012.
[28] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," in *Noise Reduction in Speech Processing*. Berlin, Germany: Springer, 2009, pp. 1–4.
[29] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control*. Hoboken, NJ, USA: Wiley, 2015.
[30] G. P. Zhang, "Time series forecasting using a hybrid ARIMA and neural network model," *Neurocomputing*, vol. 50, pp. 159–175, Jan. 2003.
[31] M. Dcosta, D. Shastri, R. Vilalta, J. K. Burgoon, and I. Pavlidis, "Perinasal indicators of deceptive behavior," in *Proc. 11th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, vol. 1, May 2015, pp. 1–8.
[32] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 1995, vol. 14, no. 2, pp. 1137–1145.
[33] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.
[34] E. T. Solovey, M. Zec, E. A. Perez, B. Reimer, and B. Mehler, "Classifying driver workload using physiological and driving performance data: Two field studies," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.* New York, NY, USA: ACM, Apr. 2014, pp. 4057–4066.
[35] J. Healey and R. Picard, "SmartCar: Detecting driver stress," in *Proc. 15th Int. Conf. Pattern Recognit.*, vol. 4, Sep. 2000, pp. 218–221.
[36] M. Miyaji, H. Kawanaka, and K. Oguri, "Driver's cognitive distraction detection using physiological features by the AdaBoost," in *Proc. 12th Int. IEEE Conf. Intell. Transp. Syst.*, Oct. 2009, pp. 1–6.
[37] D. A. Johnson and M. M. Trivedi, "Driving style recognition using a smartphone as a sensor platform," in *Proc. 14th Int. IEEE Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2011, pp. 1609–1615.
[38] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[39] G. Panagopoulos, "Multi-task learning for commercial brain computer interfaces," in *Proc. 17th Int. Conf. Bioinform. Bioeng. (BIBE)*, Oct. 2017, pp. 86–93.

George Panagopoulos received the B.E. degree in informatics and telematics from the Harokopio University of Athens, Greece, in 2014. He has worked as a Research Assistant at the Computational Physiology Lab, University of Houston (UH), and also at the Software Knowledge and Engineering Lab, NCSR Demokritos, Athens. His research interests lie in machine learning and its application to bioengineering and affective computing. The work described in this article is a part of his M.S. thesis, conducted under the supervision of Prof. I. Pavlidis. He received the Best M.S. Thesis Award from the UH Department of Computer Science for this research.

Ioannis Pavlidis (S'85–M'87–SM'00) received the B.E. degree in electrical engineering from the Democritus University of Thrace, Xanthi, Greece, in 1987, the first M.S. degree in robotics from the University of London, London, U.K., in 1989, and the second M.S. and Ph.D. degrees in computer science from the University of Minnesota, in 1995 and 1996, respectively. He is currently the Eckhard-Pfeiffer Professor of computer science and the Director of the Computational Physiology Laboratory, University of Houston. His research has been supported by multiple sources, including the National Science Foundation, the Department of Defense, and Transportation Organizations. He is the author of many scientific articles on computational physiology, affective computing, and science of science. He is well-known for his work on facial signs of stress, which first appeared in *Nature* and *Lancet*.