Understanding Driving Distractions: A Multimodal Analysis on Distraction Characterization

Michalis Papakostas
mpapakos@umich.edu
Computer Science & Engineering
University of Michigan

Kais Riani
kriani@umich.edu
Computer & Information Science
University of Michigan

Andrew Gasiorowski
abgasior@umich.edu
Computer & Information Science
University of Michigan

Yan Sun
yansu@umich.edu
Computer Science & Engineering
University of Michigan

Mohamed Abouelenien
zmohamed@umich.edu
Computer & Information Science
University of Michigan

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

Mihai Burzo
mburzo@umich.edu
Computer Science, Engineering,
Physics
University of Michigan

ABSTRACT

Distracted driving is a leading cause of accidents worldwide. The tasks of distraction detection and recognition have been traditionally addressed as computer vision problems. However, distracted behaviors are not always expressed in a visually observable way. In this work, we introduce a novel multimodal dataset of distracted driver behaviors, consisting of data collected using twelve information channels coming from visual, acoustic, near-infrared, thermal, physiological and linguistic modalities. The data were collected from 45 subjects while being exposed to four different distractions (three cognitive and one physical). For the purposes of this paper, we experiment with visual and physiological information and explore the potential of multimodal modeling for distraction recognition. In addition, we analyze the value of different modalities by identifying specific visual and physiological groups of features that contribute the most to distraction characterization. Our results highlight the advantage of multimodal representations and reveal valuable insights for the role played by the two modalities on identifying different types of driving distractions.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in ubiquitous and mobile computing**; • **Social and professional topics → User characteristics**; • **Information systems → Multimedia and multimodal retrieval**.

KEYWORDS

distracted driving, machine learning, physiological signal processing, action unit analysis, multimodal interaction, multimodal datasets

ACM Reference Format:
Michalis Papakostas, Kais Riani, Andrew Gasiorowski, Yan Sun, Mohamed Abouelenien, Rada Mihalcea, and Mihai Burzo. 2021. Understanding Driving Distractions: A Multimodal Analysis on Distraction Characterization. In *26th International Conference on Intelligent User Interfaces (IUI '21), April 14–17, 2021, College Station, TX, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3397481.3450635

1 INTRODUCTION

Road traffic accidents have increasingly become a worldwide leading cause of death and injuries. According to the Centers for Disease Control and Prevention (CDC) and the World Health Organisation (WHO), every year traffic accidents claim the lives of 1.35 million people around the world, resulting in almost 3,700 road casualties daily, which involve cars, buses, motorcycles, bicycles, trucks, and/or pedestrians [18]. While having a devastating societal impact, road accidents are highly correlated with severe financial losses as well. CDC reports that in just one year (2013), the total lifetime medical and work loss costs associated with fatal and non-fatal road injuries in the United States was estimated at 154.33 billion dollars, while 37% of the costs associated with unintentional injury deaths in general during the same year were directly related to transportation accidents [5, 19].

One of the most common causes of road accidents is distracted driving. Based on the National Highway Traffic Safety Administration (NHTSA), over the span of one year (2018), 2,800 lives were lost in US road accidents due to distracted driving and more than 400,000 thousand people were injured [20]. NHTSA defines distracted driving as any activity that diverts attention from driving, including talking or texting on the phone, eating and/or drinking, talking to people in your vehicle, fiddling with the stereo, entertainment or navigation system or anything else that takes driver's attention

IUI '21, April 14–17, 2021, College Station, TX, USA
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8017-1/21/04…$15.00
https://doi.org/10.1145/3397481.3450635

away from the task of safe driving. According to the same source, texting is the most alarming distraction. Sending or reading a text takes the driver's eyes off the road for a minimum of 5 seconds. At 55 mph, this is the same as driving the length of an entire football field with the eyes closed.

NHTSA and CDC classify driver distractions into three major categories that occupy different types of driver's mental and motor capabilities [17]: *Visual* — taking your eyes off the road; *Manual* — taking your hands of the wheel; and *Cognitive* — taking your mind off what you are doing. These distraction categories may of course overlap and coexist in many types of driving distractions.

Motivated by this previous foundational work, this paper targets the following three research questions:

(1) **How do different distractions affect driver's behavior?** We propose a novel dataset towards understanding distracted and drowsy driving. The dataset covers a group of different driving distractors and is designed with a special focus to induce different aspects of cognitive inattention motivated by variant affective stimuli.
(2) **How do different visual and physiological modalities perform with respect to capturing distracted behavior?** We perform an in-depth evaluation of different scenarios and we identify the strengths and weaknesses of each modality towards a) detecting and b) recognizing physical and cognitive distractions.
(3) **What are the most important features when detecting distracted behavior?** We perform a modality-based feature analysis on the different trained models and highlight the most informative features in each information channel.

The goal of this research is to gain insights into how distractions affect behavior. This is realized by exposing the participants to different cognitive distractions induced by affective stimuli and identifying behavioral and physiological features that can best characterize those behavioral changes.

2 RELATED WORK

The vast majority of past research focused on computer-vision based approaches to characterize distracted driving. The work by Mbouna et al. [14] in 2013 used a set of facial and head related features to tackle the problem. Eye-state monitoring and head-pose patterns were tracked overtime to classify between alert vs non-alert. This work highlighted the very rich information that can be extracted from the head and eye regions and showed its great potential towards understanding distracted behavior. The method proposed in 2015 by Liu et al. [13] tried to address the problem by acknowledging and targeting a common issue across many machine learning applications; the lack of labeled data. The research team proposed a semi-supervised method that, similar to works of the past, utilized eye and head movements to detected distractions based on both labeled and unlabeled data. In more recent works, deep-learning methods have been evaluated on similar experimental setups. The works proposed in 2019 by Kose et al. [10] and Rao et al. [25], utilized convolutional neural networks to classify video segments into 10 target classes using the dataset proposed by Abouelnaga et al. [1]. These two papers were likely the first to go beyond distraction detection to distraction recognition. However, their methods were highly dependent on discriminating physical distractors by targeting labels such as "reaching behind" or "talking on phone with the right hand", thus being very limited to other kinds of passive distractors that relate to anxiety, frustration or even verbal interaction.

In the 2014 study by Solovey et al. [32], results showed that working with physiological data alone can provide high quality information regarding driver's cognitive workload; a mental state which is highly correlated with distracted behavior. Most recently in 2020, McDonald et al. [15] discussed the advantage of ensemble learners to model driver behavior and classify distractors based on physiological markers. Overall though, physiological data have been explored in further depth only during the recent past and usually in combination with other information signals such as eye-lid movements or vehicular-based feedback signals, showing very promising results and highlighting new research directions [22, 36].

While several recent papers have tried to study the fluctuations of stress during driving [34, 35], very few have focused explicitly on how different common driving distractors affect specific physiological and behavioral reactions [37], and even fewer have explored the potential of multimodal data for such purposes [3, 26].

This paper attempts to fill some of the gaps not yet filled by previous studies. Firstly, we explore the problem of distraction detection and next, we address the task of distraction recognition in an effort to understand how different distractions can be discriminated. Our evaluation goes beyond physical distractions and aims to discriminate between distractions that involve different types of cognitive effort, such as listening and commenting on emotionally intriguing radio recordings or interacting with a faulty GPS that can cause frustration and mild levels of anxiety. Recognising different types of distractions can also lead to more personalised driving assistants, a utility that becomes more and more popular in modern vehicles. Through this analysis we aim to identify features that could be potentially evaluated in other related applications, such as education, training and other task-oriented domains [23]. Secondly, our results emphasize the advantage of multimodal distraction recognition approaches and provide valuable insights for further research on distraction characterization in driving and beyond.

3 HOW DO DIFFERENT DISTRACTIONS AFFECT DRIVER'S BEHAVIOR? - THE DATASET

We introduce a novel multimodal dataset that has been specifically developed for the purposes of understanding distracted and drowsy driving. The dataset was collected under a simulated environment using twelve different information signals on 45 subjects of varying ethnicity. Overall, the dataset consists of 30 males and 15 females, all between 20 and 33 years old. Figure 1 illustrates the experimental setup environment.

3.1 Experimental Procedure

For each participant, we held two recordings in a simulated environment. One recording took place in the morning, usually sometime from 8am to 11am, and the second recording happened during the afternoon/evening, between 4pm to 8pm. We asked all participants to schedule the morning recording as the first task in their daily

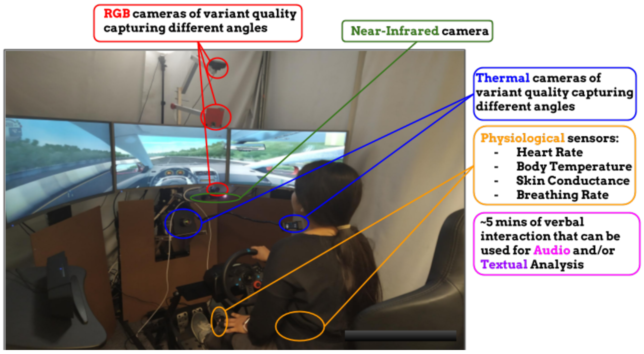
RGB cameras of variant quality capturing different angles
Near-Infrared camera
Thermal cameras of variant quality capturing different angles
Physiological sensors:
- Heart Rate
- Body Temperature
- Skin Conductance
- Breathing Rate
~5 mins of verbal interaction that can be used for Audio and/or Textual Analysis

Figure 1: The data collection experimental setup

routines so that they are as less drowsy as possible. Next, participants were asked to attend the afternoon recordings later in the day, usually before going home, and were specifically instructed not to nap in that day from the time they woke up until the time of the recording. Our assumption was that at different times of day we could capture variant levels of alertness and biological rhythms. The two recordings did not have to happen in the same day or in any specific order. Each recording lasted on average 45 minutes and consisted of three different sub-recordings; 'baseline', 'free-driving' and 'distractions'. During each session, subjects had to drive both on highways and in a city-like environment.

The 'baseline' recording consisted of two sub-parts; the 'base part' and the 'eye-tracking' part. In the 'base part' participants were asked to sit still, breath naturally and stare at the middle of the central monitor for 2.5 minutes. For the 'eye-tracking' part, subjects were shown a pre-recorded video with a target changing its position every few seconds. Participants were asked to follow the target with their gaze while acting naturally. This part lasted another 2.5 minutes.

During the 'free-driving' recording, participants had to drive uninterrupted for approximately 15 minutes. Before the beginning of each 'free-driving' recording and after explaining the basic operation controls, we gave participants a chance to drive for a few minutes so they can familiarize themselves with the simulator. To minimize the biases introduced by the relatively unfamiliar virtual-driving setup, for the purposes of this paper we used only five minute long data segments, extracted from the last seven minutes of the free-driving recording, when subjects were already used to the driving simulator.

The last part was the 'distractions' recording. This recording consisted of four different sub-parts that simulated different types of common driving distractors. The largest portion of the analysis discussed in this work has been conducted on the data collected during this part. Bellow we describe the four different distractors that participants were exposed to during each recording session.

- **D1 - Texting**. Participants were asked to type a short text message on their personal mobile device. The text was a predefined 8-word message and was dictated to the participant by the experiment supervisor on the fly. By using predefined texts we aimed to minimize the impact of cognitive effort that subjects had to put when texting and focus more on the physical disengagement from driving. Nonetheless, texting combines all three distraction classes defined by NHTSA and the CDC, which are Manual, Visual and Cognitive (see Section 1). The mobile device was placed on an adjustable holder on the right side of the steering wheel and participants had the freedom to adjust the positioning of the holder at will, so that it fits their personal preferences, thus simulating a real-car setup as accurately as possible.
- **D2 - N-Back Test**. The second distractor was the N-Back test. This distractor aimed to challenge exclusively the Cognitive capabilities of the subjects while driving. N-Back is a cognitive task extensively applied in phsycology and cognitive neuroscience, designed to measure working memory [9]. For this distractor, participants were presented with a sequence of letters, and were asked to indicate when the current letter matched the one from n steps earlier in the sequence. For our experiments we set N=1 and deployed an auditory version of the task where subjects had to listen to a prerecorded sequence of 50 letters.
- **D3 - Listening to the Radio**. For this distractor, participants were asked to listen to a pre-recorded audio from the news and then comment about what they just heard by expressing their personal thoughts. As with the N-Back Test, this distractor challenges mainly the cognitive capabilities of the participant when driving but with one major difference. In contrast to the neutral nature of the previous distractor here the recordings were emotionally provocative hence, motivating an affective response from the side of the subject. In particular, the two recordings used as stimuli for this part were related to a) a potential active shooter event that took place in the greater Detroit area and b) reporting from a fatal road accident scene which took place in the area of Chicago. These choices were made to help the users relate better to the events described in the recordings.
- **D4 - GPS Interaction**. At this step we asked participants to find a specific destination on a 'GPS' through verbal interaction. The goal of this distractor was to induce confusion and frustration to the participant; emotions that people are likely to experience when driving, either by interacting with similar 'smart' systems or through the engagement with other passengers or drivers on the road. In this case, the 'GPS' was operated by a member of the research stuff in the background providing missleading answers to the participant and repeating mostly useless information until the desired answer was provided.

What was most surprising about this section was that despite the fact that we were expecting this to be mainly a cognitive/emotional challenge, we empirically observed that very often subjects tended to take their visual attention from the driving task and repeat (often quite loudly) their commands while looking towards the direction of the speaker. Even though the scenario tested here is purely experimental and no final conclusions can be made, this observation offers a valuable insight about the general driver behavior and reaction patterns on various distractions.

Once the participants started driving they would not stop until the end of the recording. Thus, they did not experience any interruptions when switching from the 'free-driving' to the 'distractions' parts. For each of the distractors we had two similar alternatives, which we randomly switched between morning and afternoon recordings making sure that each subject would be exposed to a different stimuli each time they participated.

3.2 Modality Description

During each recording the following visual, acoustic, near-IR, thermal, physiological and linguistic modalities were recorded:

1. Top-view RGB camera from Logitech, recording at 30 fps.
2. Face closeup RGB camera from Raspberry, running on a Raspberry-Pi, recording at 25 fps.
3. Face closeup RGB camera from IDS, capturing data at 20 fps.
4. Near-Infrared close-up camera from IDS, capturing data at 20fps.
5. Low quality thermal camera from Flir, capturing the face of the subject with a small angle from the center, at an average of 7 fps.
6. High quality thermal camera from Flir, capturing the subject's face at 100 fps.
7. Four physiological sensors from Thought Technology Ltd., 3 of them attached on the non-dominant hand of the subject and one on the torso, measuring the following information: a) *Blood Volume Pulse* (BVP), b) *Skin Temperature*, c) *Skin Conductance* and d) *Respiration*.
   Raw data were captured with a sampling frequency of 2048 Hz. To avoid excessive amounts of noise in the physiological data, we asked subjects to drive mostly using only their dominant hand and use their other hand only if needed.
8. Audio was recorded during the 'Listening to Radio' and the 'GPS Interaction' distractors, where subjects had to provide verbal feedback.
9. Transcriptions of the audio recordings are also available.

We also recorded the driver's simulation run. For the purposes of this paper we focus exclusively on the data captured from the sensors in (3) and (7), i.e., the close-up RGB video recorded with the IDS camera and the four physiological indicators.

4 METHODOLOGY

In this work, we try to address two different problems. Distraction detection, i.e., characterize the subject as distracted or not and distraction recognition, i.e., identify the type of distraction that the subject is involved in. For each task we perform experiments using modalities individually, while for distraction recognition we explore the potential of multimodal processing as well, by testing both early and late fusion of the two modalities. In the following paragraphs we discuss the pre-processing and feature extraction steps for each of the modeling approaches.

4.1 RGB close-up Image Processing

Inspired by the promising results of past research (see Section 2), we analyze features extracted from the face and head regions using the Openface library [2]. Openface estimates a rich set of facial and head positioning features based on a Constrained Local Model that consists of two main components; a Point Distribution Model that is responsible for modeling the shape of a face and a group of local detectors responsible to evaluate the probability of a landmark being aligned at a particular pixel location [29]. Output features provided by Openface include head pose and eye gaze information, facial landmark coordinates and action unit (AU) presence as well as intensity values. We experiment with both individual and combinations of those features and we conclude that AU intensity values were the ones encapsulating the richest amount of information for our scope.

To describe AUs we first need to introduce the Facial Action Coding System (FACS). FACS is a framework designed to group facial movements based on their appearance on the human face. This grouping depends on slight instant changes in facial appearance caused by individual face muscles. AUs are the individual units used by FACS to code complex facial expressions. Thus, AUs can be seen as a mid-level representation of facial expressions, providing higher level of information than just a group of facial landmarks but being much more descriptive than an affect-based classification or regression model [33].

Openface provides AU intensity in the form of a continuous variable for 17 different AUs. Intensity values may range from zero (AU is not present) to five (maximum intensity). The AUs monitored by Openface can be seen in Figure 6.

We compute intensity values for all 17 AUs for every frame in our video data. Following that a sliding window technique is applied to the sequence of frames. For our experiments a two second window with 50% overlap was used. Hyper-parameters were tuned through an exhaustive grid search approach. For every window we extract the following features describing the distribution of AU intensities within a window; minimum and maximum values, average, variance, skeweness and kyrtosis. At the end of this process each window is summarized to a $17 \times 6 = 102$ feature vector.

4.2 Physiological Data Processing

As discussed in Section 3.2 we collect four physiological indicators with a sampling rate of 2048 Hz. For each of these signals, domain-specific statistical features are extracted using the BioGraph Infiniti data processing platform [16].Every feature value computed over a group of raw measurements, is also used as a padding value until the next computation, so that the final output matches the sampling rate of the raw data.

In total 73 domain-specific features are extracted through BioGraph in the form of time-series from all four raw data streams. From these 73 features 49 are related to BVP, six to Skin Temperature, eight are extracted from Respiration, six are bi-products of Skin Conductance and four features are statistics correlating heart rate with breathes per minute.

The BVP related features, which have the lion's share in the final data representation are extracted from both the temporal and frequency domains of the raw signal. In particular, there are ten features describing the statistical behavior of the inter-beat intervals of the BVP signal, ie. distance between BVP peaks. Moreover, twelve features are related to heart rate (HR) and heart rate variability (HRV), describing temporal statistics of HR and frequency related information for HRV such as low to high frequency ratio, peak frequency and others. Additionally, 24 features are computed to describe the spectral power statistics of different frequency bands on the BVP signal by grouping the frequencies into three frequency groups, very-low (<0.04 Hz), low(0.04-0.15 Hz) and high frequencies (0.15-0.4 Hz). For each frequency band 8 power related statistics are calculated describing the total power of that frequency band, its mean and standard deviation and their corresponding percentages with respect to the complete signal at the time that the measurement is taken. Lastly, three features are extracted to describe the behavior of the amplitude of the raw BVP signal at each timestamp.

The remaining 24 features extracted from the Skin Temperature, Respiration and Skin Conductance streams are statistics describing exclusively the temporal behavior of each of the signals and the correlation of individual measurements with respect to their maximum and minimum values. Adding the four raw data measurements to the 73 features described above, we end up at each timestamp with a set of 77 domain-specific "core features" describing the physiological state of the participant.

Next we segment the 77 information streams using again a sliding window approach with a window size of four seconds and a 50% overlap. As before, hyper-parameters are tuned using an exhaustive grid search approach. At every temporal window we compute for every feature the same six statistics mentioned in Section 4.1 plus the zero-crossing rate [27]. Zero crossing rate indicates the rate of sign-changes of a signal during the duration of a particular frame and is often used for audio and physiological signal modeling tasks.

From each of the 77 information streams, we thus compute seven statistics resulting in a 539 features set. In addition, we compute the first order difference between the current and the previous frame, eventually resulting in a final feature vector of 1078 features representing four seconds of physiological measurements.

4.3 Modality Fusion

For our distraction recognition experiments we apply multimodal processing using a late and an early fusion approach. For both methods we perform one prediction every two seconds taking into account data captured over a period of four seconds. Given the modality-based windowing approaches described above there are about three feature vectors coming from the visual modality corresponding to every feature vector coming from the physiological sensors.

In the late fusion approach, we classify each modality independently and then merge the probabilities of the two different models to assign the final prediction. To do so, we first average the class probabilities produced by the visual classifier for every three consecutive samples, and then sum the predicted class probabilities provided by each modality-model. The label assigned is the one with the maximum final value.

For the early fusion approach the process is very similar with the main difference being that we compute an average visual feature vector for every three consecutive visual samples. Then the averaged vector is being concatenated with the corresponding physiological sample, resulting to a final representation of 1078+102=1180 features. This representation is then normalized using a standard normalization and passed through the classification algorithm.

4.4 Classification

To evaluate the effectiveness and robustness of the proposed modality representations, we compute the classification performance using different types of classifiers. Here we report results for two classifiers:

(1) Ensemble Voting: A Random-Forest (RF) classifier using 100 Decision Trees, with a maximum depth of 100 features per tree. We use entropy as a metric to ensure maximum information gain at each node [12].
(2) Ensemble Boosting: A Gradient Boosting (GB) classifier that estimates a final set of weights for each sample based on an iterative process. For our experiments we use 100 weak estimators [6].

We also experimented with an SVM classifier with a linear and an RBF kernel but the results were always comparable or worse than the other two alternatives. In addition, an important benefit of the ensemble classifiers compared to SVM is the interpretability of results. These observations are also in line with other related studies [15]. In Section 5.2 we decompose the different ensemble models to better understand feature importance and contribution to the final results.

5 EXPERIMENTAL FINDINGS

We conduct three types of evaluation experiments. Initially we target the traditional problem of distracted versus non-distracted driving. Next we look deeper into the distractions and instigate two novel experimental setups towards better understanding the nature of different distractors. First we address the binary task of discriminating between physical (D1-Texting) and mental (D2-NBack, C3-Radio and D4-GPS) distractors. Second, we repeat the experiment by considering each distraction as an individual class. We approach all problems using each modality independently and for the distraction recognition tasks we also report results following the multimodal approaches discussed in Section 4.3. After the quantitative analysis provided by the classification results we discuss a qualitative evaluation that aims to identify features that contributed the most.

5.1 How do Different Visual and Physiological Modalities Perform with respect to Capturing Distracted Behavior?

For all our results we perform a leave one subject out cross validation and report performance in terms of average F1. Given the variant complexity and nature of each problem addressed in this paper we believe that average F1 offers the ground to produce comparable and balanced results that avoid data distribution biases affecting other metrics such as accuracy. For the models with maximum F1 on the task, we visualize the averaged confusion matrices and evaluate deeper by discussing recall performances of individual classes. In all the following tables, bold values correspond to the best result in each experimental setup.

Finally we run each experiment following three different modeling approaches:

(1) User Independent: We used all the data from 44 users for training and the remaining user for testing. We repeated the

process 45 times until all users were used as a test set. At the end the results of all 45 models were averaged.

(2) User Dependent: We used all the data from 44 users for training. For the 45^{th} user we included one of his recordings (morning or afternoon) in the training data and the remaining recording was used for testing. We repeated the process 90 times until all users were used as a test set. At the end the results of all 90 models were averaged.

(3) User Exclusive: For each user we used one of their recordings (morning or afternoon) for training and their remaining recording for testing. No data from other users were included in the training or testing set in this case. We repeated the process 90 times (2 times for each of the 45 users). At the end the results of all 90 models were averaged.

For all our experiments we compare with two baselines. First we show results based on a weighted classifier which always led to maximum average F1. For this baseline the chance of assigning a label to a sample is equal to the percentage of samples available for each class in the training dataset. Since the baseline predictions were weighted based on the class probabilities the final average F1 (computed across all folds) always converged to $\frac{1}{\#classes}$. We refer to that baseline as "Balanced". As an additional baseline we report average performance when assigning the same label to all test samples. This can be considered as a more balanced version of the majority class classifier since the final result takes into account performance across all the individual classes. We refer to this classifier as "Single label". The reported results were evaluated for significance using a non-parametric Wilcoxon test showing always strong evidence of difference against baseline with p values ranging from 1^{-14} to 0.03.

Given the different experimental setups tested, the exact amount of training and testing data used in each fold of each experiment varies. Table 1 shows the total duration of data used for our experiments under each recording segment.

5.1.1 ***Distraction Detection***. For this experiment we first segment 5-minute long recordings coming from the last seven minutes of the free-driving recording part (see Section 3.1). The distracted class contains samples collected during the distraction recording parts and are a mix of all four distractors. Distraction samples correspond to 60% of the samples while Free-driving data occupy the remaining 40%. Table 2 shows the classification results, while Figure 2 illustrates the confusion matrices for the best results using each modality.

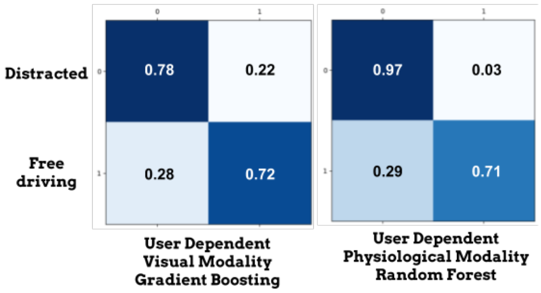

Figure 2: Confusion matrices on distracted VS non-distracted driving classification for the best results of Table 2

The results of Table 2 indicate that tuning the model with user-specific data enhances F1 performance compared to just training on generally observable behavioral patterns. In addition, the matrices of Figure 2 reveal that the physiological model significantly outperformed the visual one in terms of recall performance for the 'Distracted' class. The physiological model showed an absolute improvement of 18%, by correctly identifying 97% of the distracted samples. The two models have very comparable performance on detection of 'non-distracted' samples. In general, the 'User Dependent' model of the physiological sensors trained on an RF classifier, offered the best results with 87% average F1 and 97% and 71% average recall for the distraction and non-distraction classes accordingly. This highlights the robustness of the physiological modalities on detecting patterns of inattention that are not visually observable, as head and face based features are.

5.1.2 ***Distraction Recognition***. In this scenario, we try to identify different distractions based on their nature. For the binary problem, i.e., physical versus mental distractions, the latter represent the dominant class with 71% of the total number of samples. For the 4-class problem 29% of the data belongs to distractor D1, 20% to D2, 32% comes from D3, and 19% from D4. Tables 3 and 4 show the results on each experiment and Figures 3 and 4 their corresponding confusion matrices. Lastly, Table 5 and Figure 5 present results based on early and late multimodal fusion.

Overall, all modeling approaches (unimodal in Table 3 and multimodal in Table 5) perform very well on discriminating physical from mental distractors. The physical activity demanded by the subject to text, generates motion patterns that both modalities can easily pick-up. What is interesting is the very high performance observed by the visual modality in the 'User Exclusive' experiment, shown in Table 3, where the available training data were very limited compared to the other two experimental-setups. This highlights the value of the vision-based method on detecting physical distractions. However, AU-based modeling seems unable to depict significant behavioral differences across subjects, which translates to the minor increase in performance in the 'User Independent' and 'Use Dependent' approaches compared to the corresponding improvements observed between the different physiological models.

On the other hand, when we increase the resolution of the targeted classes, physiological sensors are much more robust on discriminating between cognitive distractors of different stimuli. The lack of significant motion activity makes the visual sensor a weaker descriptor and in general less flexible to compete. This can be confirmed by both Table 4 and Figure 4.

Nonetheless, in both distraction detection tasks, multimodal modeling provides the best results almost in all scenarios. These results are shown on Table 5 and Figure 5. While, both fusion mechanisms show improvements over the unimodal methods, late fusion provids the best performance reaching a 94% average F1 on the 2-class problem and 67% on the 4-class task. Comparing the confusion matrices shown in Figure 5 to their unimodal counters, we observe important improvements in some cases or stable performances in others. The only case where a unimodal method outperforms the multimodal approach is in the recall of the D3 -'Radio' distraction in the 4-class problem, where the multimodal approach recorded a 61% average F1 compared to the 81% recorded by the physiological

Table 1: Total duration of available data under each recording segment

	Recording Segment				
	Freedriving	Texting Physical	NBack Cognitive Neutral	Radio Cognitive Emotional	GPS Cognitive Frustration
#Data (hours)	~7.4	~3.1	~2.2	~3.4	~2

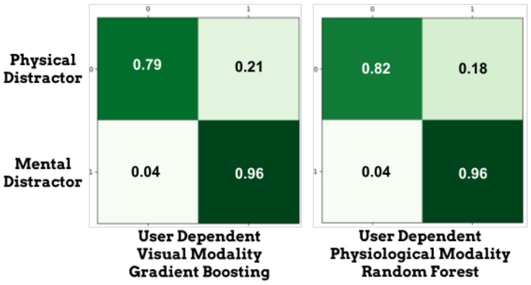

Figure 3: Confusion matrices on distraction recognition as a 2-class problem for the best results of Table 3

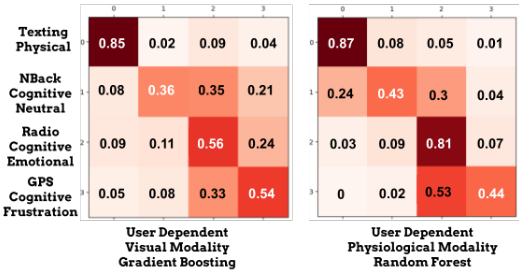

Figure 4: Confusion matrices on distraction recognition as a 4-class problem for the best results of Table 4

Table 2: Results on distracted VS non-distracted driving classification with respect to average F1.

	Baseline		Visual		Physiological	
	Single Label	Balanced	RF	GB	RF	GB
User Independent	0.32	0.5	0.69	**0.7**	**0.86**	0.84
User Dependent	0.32	0.5	0.73	**0.75**	**0.87**	0.84
User Exclusive	0.32	0.5	**0.68**	0.65	**0.53**	**0.53**

Table 3: Results distraction recognition as a 2-class problem with respect to average F1.

	Baseline		Visual		Physiological	
	Single Label	Balanced	RF	GB	RF	GB
User Independent	0.32	0.5	0.84	**0.86**	**0.89**	0.88
User Dependent	0.32	0.5	0.85	**0.88**	**0.90**	0.88
User Exclusive	0.32	0.5	**0.85**	0.79	**0.63**	**0.59**

Table 4: Results on distraction recognition as a 4-class problem with respect to average F1

	Baseline		Visual		Physiological	
	Single Label	Balanced	RF	GB	RF	GB
User Independent	0.1	0.25	0.47	**0.53**	**0.64**	0.63
User Dependent	0.1	0.25	0.5	**0.58**	**0.65**	0.64
User Exclusive	0.1	0.25	**0.51**	0.47	**0.31**	**0.31**

Table 5: Results on distraction recognition using multimodal learning.

		Early Fusion		Late Fusion
		RF	GB	
2-class	User Independent	0.9	0.87	**0.92**
	User Dependent	0.91	0.89	**0.94**
	User Exclusive	0.63	0.58	**0.84**
4-class	User Independent	**0.63**	**0.63**	0.62
	User Dependent	0.64	0.66	**0.67**
	User Exclusive	0.31	0.32	**0.46**

data in Figure 3. However, the result is still better compared to the 56% recall reported by the visual modality in the same figure.

Lastly, it is important to mention the improved performance on the 4-class problem by the multimodal model shown in Figure 5. Special attention has to be given to the models' ability to differentiate between cognitive distractors based on the nature of their stimuli. The classifier shows improved performance over both unimodal models for the three cognitive-classes with an average absolute improvement of 0.14 recall units over the visual modality and 0.06 recall units over the physiological modality (maximum recall is 1). Further statistical analysis for significance testing on the 4-class problem partially proves those observations showing a clear difference between the multimodal and the visual models, with p values bellow 6×10^{-4}. The case is not the same when comparing to the physiological models where p values range around 0.1, thus highlighting the great impact of the physiological modality on the final result.

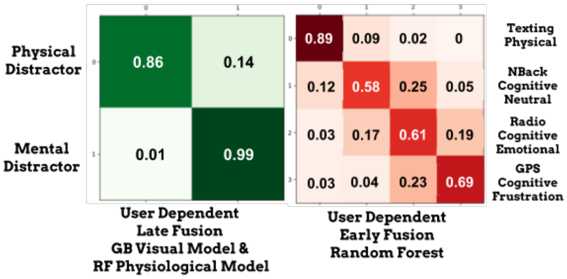

Figure 5: Confusion matrices on distraction recognition for the best results of Table 5. Left: Best results for the binary approach, Right: Best results on the 4-class approach

5.2 What are the most Important Features when Detecting Distracted Behavior?

Figure 6 illustrates the intensities of different AUs with respect to the four distractors. There are some clear trends in several cases such as in AU4 and AU14 which, seem to be more present in the D1-'Texting' distractor. Similarly AU6, AU15 and AU26 seem to be quite active during distractor D4 -'GPS Interaction', which was designed to induce communication dissonance and frustration to the subject. AU15, AU17 and AU25 are also present during distractor D3 - "Listening to the Radio". There are no clear trends between AU intensities and the NBack - neutral distractor.

Next we look into the importance of different statistical features extracted from each modality. Feature importance is calculated as

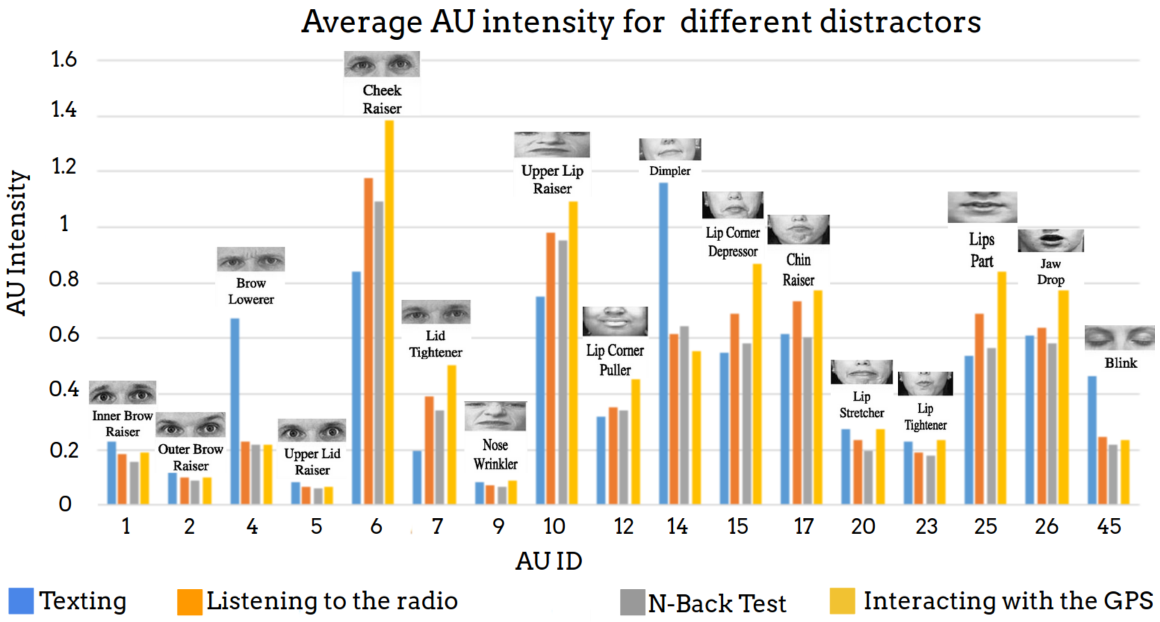

Figure 6: AU Intensities per distractor. AU ID numbers are defined by FACS.

the increase in information gain at each node or in other words the decrease of information entropy caused by each feature. The higher the value the more important the feature.

We use Python's Scikit-learn implementation for estimating feature importance [24]. For visualization purposes, we average feature importance values across all models trained on all three unimodal classification tasks given a classifier and a modality. For the visual modality we use as a reference the GB classifiers and for the physiological the RF models, as they respectively showed best performance on each corresponding modality . Figures 7.a, 7.b show feature importance values for each feature, ie. 102 visual and 1078 physiological features (see Sections 4.1,4.2). Table 6 presents the top#5 performing "core features" from each signal. By "core features" we refer to the initial intensities for the 17 AUs and the 77 domain-specific physiological features before extracting window-based statistics.

Table 6: Top 5 features of each modality based on feature importance

	AU	Physiological
#1	AU05	BVP IBI pNN Intervals (%)
#2	AU10	BVP IBI pNN Intervals
#3	AU06	BVP HF % power mean
#4	AU23	BVP LF % power mean
#5	AU01	BVP IBI NN Intervals

It is worth observing that across the visual features (Figure 7.a) some of the most informative ones come from AUs that show overall low intensity levels when judging from Figure 7, in particular AU1, AU5 and AU23. This indicates that differences in AU intensity that seem minor to the naked eye may be crucial towards identifying distracted behavior. For the physiological sensors (Figure 7.b), BVP related features seem to account the most for the good results offered by the modality. However, as seen in Figure 7.b, all physiological indicators contributed to the final results despite the fact that the vast majority of features were related to BVP. The top#5 most informative physiological measurements are presented bellow:

- BVP IBI pNN Intervals (%): the percentage of successive intervals that differ by more than 50 ms
- BVP IBI pNN Intervals: the number of successive intervals that differ by more than 50 ms
- BVP HF power mean:the mean of power in the high frequencies
- BVP LF power mean: the mean of power in the low frequencies
- BVP IBI NN Intervals: interval between two normal heartbeats

pNN features are known to be highly correlated with sympathetic and parasympathetic modulation of the nervous system [7, 28]. The sympathetic nervous system is responsible to release hormones that accelerate the heart rate, while the parasympathetic has the opposing role. Factors as stress, caffeine, and excitement may temporarily accelerate heart rate stimulated by the sympathetic system[11].

To emphasize the significance of the features reported on Table 6 and get a deeper understanding of their impact in the overall decision making we repeat all the best performing experiments (User-Dependent scheme) using only those top#5 variables from each modality. We show our results in Figures 7.c, 7.d.

Interestingly enough, almost in all cases only a minor decrease in performance is observed, highlighting the increased performance of the selected features over the complete set. The deepest decrease in performance can be found on the visual modality for the distraction

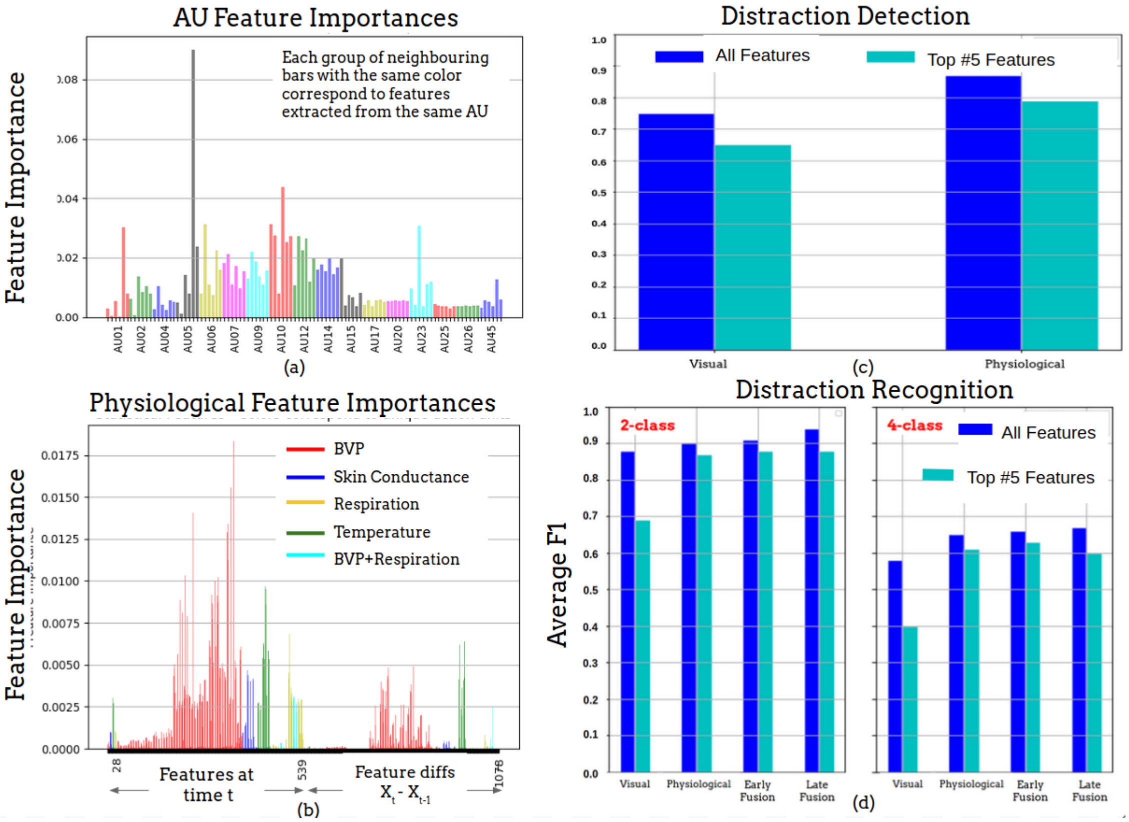

Figure 7: (a), (b): Feature importance for each modality based on information gain. (a): Visual Features, (b): Physiological Features. (c), (d): Performance comparison between best models (User-Dependent scheme) trained on the top 5 features of each modality versus all the available features. Graph (c) corresponds to Distraction Detection, graphs in (d) to Distraction Recognition ((d)-left as 2-class, (d)-right as 4-class)

recognition task. In both the 2-class and 4-class scenarios, visual-only performance had an absolute decrease of more that 15% in terms of average F1, signifying again how volatile the visual features can be on this task when used as a standalone modality.

On the other hand the gains of multimodal approaches are now less obvious. We believe that in early-fusion the models could not capture important relations between the different variables due to the limited amount of available features. This property, can also become the advantage of such models when richer feature sets are available, as they have the ability to learn more insightful feature relations across the different modalities. In the case of late-fusion, the drop in performance is expected and can be justified by the decrease in performance observed in the unimodal components of the fusion algorithm.

6 CONCLUSIONS

In this paper, we presented a data-driven, machine-learning-based analysis for the tasks of driving distraction detection and recognition through visual and physiological sensors. Despite the experimental nature of our setup, there is substantial research evidence to support the direct application and integration of our methods in modern vehicles [4, 8, 21, 30, 31].

Our work highlights the trade-offs that each of the explored modalities brings to the table. In addition, it provides a fine-grained list of modality specific features which are crucial towards detecting and characterizing common physical and cognitive driving distractions.

Revisiting the research questions defined in Section 1, the contribution of this paper can be summarized as follows:

How do different distractions affect driver's behavior? We proposed a novel dataset to explore this question. The dataset includes twelve different modalities and was designed to address drowsy and distracted driving, with a focus on cognitive distractions. Our initial experiments proved the value of this resource in identifying behavioral features associated with distracted behavior. We aim to research this question further by investigating the additional resources presented in Section 3.2. Our findings showed that different stimuli are correlated with specific physiological and behavioral features.

How do different visual and physiological modalities perform with respect to capturing distracted behavior? Our findings indicate that the visual modality came short of characterizing cognitive inattention. Physiological signals proved to be more effective for this task and showed a more robust performance in general. On the other hand, the visual modality showed a clear advantage in detecting physical distractors even when data were very limited. However, AU-based modeling seemed to be limited in scalability as performance was not drastically affected when the number of training samples increased.

What are the most important features when detecting distracted behavior? While other features seem to contribute primarily through their absence, in terms of physiological measures, the features describing the power of spectrum on the BVP signal are by far the most effective. The rest of the signals had less of an impact on the final result even though their contribution remained notable, especially in the multimodal experiments.

To request access to the repository please contact us at zmohamed@umich.edu.

ACKNOWLEDGMENTS

This material is based in part upon work supported by the Toyota Research Institute ("TRI"). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of TRI or any other Toyota entity.

REFERENCES

[1] Yehya Abouelnaga, Hesham M Eraqi, and Mohamed N Moustafa. 2017. Real-time distracted driver posture classification. *arXiv preprint arXiv:1706.09498* (2017).

[2] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1–10.

[3] Daniela Cardone, David Perpetuini, Chiara Filippini, Edoardo Spadolini, Lorenza Mancini, Antonio Maria Chiarelli, and Arcangelo Merla. 2020. Driver Stress State Evaluation by Means of Thermal Imaging: A Supervised Machine Learning Approach Based on ECG Signal. *Applied Sciences* 10, 16 (Aug 2020), 5673. https://doi.org/10.3390/app10165673

[4] Hyun-Seung Cho, Jin-Hee Yang, Sang-Min Kim, Jeong-Whan Lee, Hwi-Kuen Kwak, Je-Wook Chae, and Joo-Hyeon Lee. 2020. Development of a Chest-Belt-Type Biosignal-Monitoring Wearable Platform System. *Journal of Electrical Engineering & Technology* 15, 4 (2020), 1847–1855.

[5] Curtis Florence, Thomas Simon, Tamara Haegerich, Feijun Luo, and Chao Zhou. 2015. Estimated lifetime medical and work-loss costs of fatal injuries—United States, 2013. *Morbidity and Mortality Weekly Report* 64, 38 (2015), 1074–1077.

[6] Jerome H Friedman. 2002. Stochastic gradient boosting. *Computational statistics & data analysis* 38, 4 (2002), 367–378.

[7] Guido Grassi, Sabrina Vailati, Giovanni Bertinieri, Gino Seravalle, Maria Luisa Stella, Raffaella Dell'Oro, and Giuseppe Mancia. 1998. Heart rate as marker of sympathetic activity. *Journal of hypertension* 16, 11 (1998), 1635–1639.

[8] David Michael Herman. 2020. Monitoring of steering wheel engagement for autonomous vehicles. US Patent App. 16/294,541.

[9] Michael J Kane, Andrew RA Conway, Timothy K Miura, and Gregory JH Colflesh. 2007. Working memory, attention control, and the N-back task: a question of construct validity. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 33, 3 (2007), 615.

[10] Neslihan Kose, Okan Kopuklu, Alexander Unnervik, and Gerhard Rigoll. 2019. Real-Time Driver State Monitoring Using a CNN Based Spatio-Temporal Approach. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, 3236–3242.

[11] Matthew N Levy. 1971. Brief reviews: sympathetic-parasympathetic interactions in the heart. *Circulation research* 29, 5 (1971), 437–445.

[12] Andy Liaw, Matthew Wiener, et al. 2002. Classification and regression by randomForest. *R news* 2, 3 (2002), 18–22.

[13] Tianchi Liu, Yan Yang, Guang-Bin Huang, Yong Kiang Yeo, and Zhiping Lin. 2015. Driver distraction detection using semi-supervised machine learning. *IEEE transactions on intelligent transportation systems* 17, 4 (2015), 1108–1120.

[14] Ralph Oyini Mbouna, Seong G Kong, and Myung-Geun Chun. 2013. Visual analysis of eye state and head pose for driver alertness monitoring. *IEEE transactions on intelligent transportation systems* 14, 3 (2013), 1462–1469.

[15] Anthony D. McDonald, Thomas K. Ferris, and Tyler A. Wiener. 2020. Classification of Driver Distraction: A Comprehensive Analysis of Feature Generation, Machine Learning, and Input Measures. *Human Factors* 62, 6 (2020), 1019–1035. https://doi.org/10.1177/0018720819856454 arXiv:https://doi.org/10.1177/0018720819856454 PMID: 31237788.

[16] H Meyers. 2010. ProComp Infiniti/BioGraph Infiniti biofeedback system (version 5.1. 2). *Montreal, QB: Thought Technology Ltd* (2010).

[17] Centers for Disease Control & Prevention (CDC) National Center for Injury Prevention & Control. 2019. Distracted Driving. https://www.cdc.gov/motorvehiclesafety/distracted_driving/index.html. [Online; accessed 13-April-2020].

[18] Centers for Disease Control & Prevention (CDC) National Center for Injury Prevention & Control. 2019. Road Traffic Injuries and Deaths—A Global Problem. https://www.cdc.gov/injury/features/global-road-safety/index.html. [Online; accessed 13-April-2020].

[19] Centers for Disease Control & Prevention (CDC) National Center for Injury Prevention & Control. 2020. Cost of Injury Data. https://www.cdc.gov/injury/wisqars/cost/index.html. [Online; accessed 13-April-2020].

[20] US Department of Transportation National Highway Traffic Safety Administration (NHTSA). 2019. Distracted Driving. https://www.nhtsa.gov/risky-driving/distracted-driving. [Online; accessed 13-April-2020].

[21] Shotaro Odate, Naohiro Sakamoto, and Yukinori Midorikawa. 2020. *Development of Electrostatic Capacity Type Steering Sensor Using Conductive Leather.* Technical Report. SAE Technical Paper.

[22] Michalis Papakostas, Kapotaksha Das, Mohamed Abouelenien, Rada Mihalcea, and Mihai Burzo. 2021. Distracted and Drowsy Driving Modeling Using Deep Physiological Representations and Multitask Learning. *Applied Sciences* 11, 1 (2021), 88.

[23] Michalis Papakostas, Akilesh Rajavenkatanarayanan, and Fillia Makedon. 2019. CogBeacon: A Multi-Modal Dataset and Data-Collection Platform for Modeling Cognitive Fatigue. *Technologies* 7, 2 (2019), 46.

[24] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 12 (2011), 2825–2830.

[25] Xuli Rao, Feng Lin, Zhide Chen, and Jiaxu Zhao. 2019. Distracted driving recognition method based on deep convolutional neural network. *Journal of Ambient Intelligence and Humanized Computing* (2019), 1–8.

[26] Kais Riani, Michalis Papakostas, Hussein Kokash, Mohamed Abouelenien, Mihai Burzo, and Rada Mihalcea. 2020. Towards Detecting Levels of Alertness in Drivers Using Multiple Modalities. In *Proceedings of the 13th ACM International Conference on PErvasive Technologies Related to Assistive Environments* (Corfu, Greece) *(PETRA '20)*. Association for Computing Machinery, New York, NY, USA, Article 12, 9 pages. https://doi.org/10.1145/3389189.3389192

[27] Fabien Ringeval, Björn Schuller, Michel Valstar, Shashank Jaiswal, Erik Marchi, Denis Lalanne, Roddy Cowie, and Maja Pantic. 2015. Av+ ec 2015: The first affect recognition challenge bridging across audio, video, and physiological data. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge.* 3–8.

[28] Sol M Rodríguez-Colón, Fan He, Edward O Bixler, Julio Fernandez-Mendoza, Alexandros N Vgontzas, Susan Calhoun, Zhi-Jie Zheng, and Duanping Liao. 2015. Sleep variability and cardiac autonomic modulation in adolescents–Penn State Child Cohort (PSCC) study. *Sleep medicine* 16, 1 (2015), 67–72.

[29] Jason M Saragih, Simon Lucey, and Jeffrey F Cohn. 2011. Deformable model fitting by regularized landmark mean-shift. *International journal of computer vision* 91, 2 (2011), 200–215.

[30] Pragya Sharma, Xiaonan Hui, Jianlin Zhou, Thomas B Conroy, and Edwin C Kan. 2020. Wearable radio-frequency sensing of respiratory rate, respiratory volume, and heart rate. *NPJ digital medicine* 3, 1 (2020), 1–10.

[31] GiriBabu Sinnapolu and Shadi Alawneh. 2020. Intelligent wearable heart rate sensor implementation for in-vehicle infotainment and assistance. *Internet of Things* 12 (2020), 100277.

[32] Erin T Solovey, Marin Zec, Enrique Abdon Garcia Perez, Bryan Reimer, and Bruce Mehler. 2014. Classifying driver workload using physiological and driving performance data: two field studies. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* 4057–4066.

[33] Y-I Tian, Takeo Kanade, and Jeffrey F Cohn. 2001. Recognizing action units for facial expression analysis. *IEEE Transactions on pattern analysis and machine intelligence* 23, 2 (2001), 97–115.

[34] R. Verma, B. Mitra, and Sandip Chakraborty. 2019. Avoiding Stress Driving: Online Trip Recommendation from Driving Behavior Prediction. *2019 IEEE International Conference on Pervasive Computing and Communications (PerCom* (2019), 1–10.

[35] K. Wang, Y. L. Murphey, Y. Zhou, X. Hu, and X. Zhang. 2019. Detection of driver stress in real-world driving environment using physiological signals. In *2019 IEEE 17th International Conference on Industrial Informatics (INDIN)*, Vol. 1. 1807–1814.

[36] Yongquan Xie, Yi L Murphey, and Dev Kochhar. 2019. Personalized Driver Workload Estimation Using Deep Neural Network Learning from Physiological and Vehicle Signals. *IEEE Transactions on Intelligent Vehicles* (2019).

[37] Sebastian Zepf, Neska El Haouij, Jinmo Lee, Asma Ghandeharioun, Javier Hernandez, and Rosalind W. Picard. 2020. Studying Personalized Just-in-Time Auditory Breathing Guides and Potential Safety Implications during Simulated Driving *(UMAP '20)*. Association for Computing Machinery, New York, NY, USA, 275–283. https://doi.org/10.1145/3340631.3394854