Detection and Recognition of Driver Distraction Using Multimodal Signals

KAPOTAKSHA DAS, MICHALIS PAPAKOSTAS, KAIS RIANI, ANDREW GASIOROWSKI, MOHAMED ABOUELENIEN, MIHAI BURZO, and RADA MIHALCEA, University of Michigan

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 performed experiments with visual, physiological, and thermal information to explore potential of multimodal modeling for distraction recognition. In addition, we analyze the value of different modalities by identifying specific visual, physiological, and thermal 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 three 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;**

Additional Key Words and Phrases: Distracted driving, machine learning, physiological signal processing, action unit analysis, thermal (keyword), multimodal interaction, multimodal datasets

ACM Reference format:
Kapotaksha Das, Michalis Papakostas, Kais Riani, Andrew Gasiorowski, Mohamed Abouelenien, Mihai Burzo, and Rada Mihalcea. 2022. Detection and Recognition of Driver Distraction Using Multimodal Signals. *ACM Trans. Interact. Intell. Syst.* 12, 4, Article 33 (December 2022), 28 pages.
https://doi.org/10.1145/3519267

The reviewing of this article was managed by special issue associate editors Tracy Hammond, Bart Knijnenburg, John O'Donovan, Paul Taele.
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.
Authors' addresses: K. Das, K. Riani, A. Gasiorowski, and M. Abouelenien, Computer & Information Science, University of Michigan; emails: {takposha, kriani, abgasior, zmohamed}@umich.edu; M. Papakostas and R. Mihalcea, Computer Science & Engineering, University of Michigan; emails: {mpapakos, mihalcea}@umich.edu; M. Burzo, Mechanical Engineering, University of Michigan; email: mburzo@umich.edu.

2160-6455/2022/12-ART33 $15.00
https://doi.org/10.1145/3519267

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 [29]. 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 [10, 30].

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 [31]. 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 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 [28]: *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 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 clues perform with respect to capturing distracted behavior?** We explore how visual cues in the form of Action Units can be modelled using machine learning in order to detect and recognize different kinds of distracted behavior.
(3) **How do different physiological signals perform with respect to capturing distracted behavior?** We explore how different physiological signals such as the heart rate, respiration rate, skin conductance and skin temperature cues can be modelled using machine learning in order to detect and recognize different kinds of distracted behavior.
(4) **Can the thermal modality, as a relatively newer approach, detect 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.
(5) **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, physiological, and thermal features that can best characterize those behavioral changes.

2 RELATED WORK

There have been several works in the past that used machine learning methods to detect distracted driving, with the vast majority of them focusing on computer-vision based approaches and facial analytics. One of the earlier papers published on the topic was the work proposed by Rongben et al. [40] in 2004, which utilized mouth deformation tracking to detect potential behaviors of risk on drivers, such as yawning or signs of conversation. Since then, multiple papers have been published that aim to tackle the same problem following similar approaches but utilizing more novel and sophisticated computational methods.

The work published in 2013 by Mbouna et al. [24] used a set of facial and head related features to target distracted driving. Eye-state monitoring and head-pose patterns were tracked overtime to classify between alert versus 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. [22] 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. [18] and Rao et al. [35] utilized convolutional neural networks to classify video segments into 10 target classes using the dataset proposed by Abouelnaga et al. [2]. 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.

An approach that is increasingly gaining the attention of related research as modern cars are being equipped with more advanced sensors is physiological based driver modeling [27, 45]. The review provided by Begum et al. [5] offered a detailed overview of the early approaches on distracted driving detection using physiological data. Since then, things have not drastically changed as the community keeps addressing the topic based on signals related to respiration, heart rate, muscle activity and visual clues. However, research has slowly shifted from understanding statistical correlations to building driver-centric behavior models based on the aforementioned signals.

In the study of 2014 conducted by Solovey et al. [48], results showed that working with physiological data alone can provide high quality information regarding a driver's cognitive workload; a mental state which is highly correlated with distracted behavior. Similarly, the work by Dobbins et al. [9] in 2018 showed that machine learning methods can be applied on physiological data towards inferring driving and task related characteristics such as driving speed or the type of the road. Taamneh et al. [49] designed a multimodal repository of simulated drivings targeting different types of distractions. Taamneh's paper investigated the problem under a user-centric multimodal perspective using a rich set of devices. However, other than offering an in-depth and insightful statistical analysis of their findings, the research team did not provide any machine-learning based results neither identified specific modality-based features related to different distractors, which are two of the main scopes of this paper. Our work offers three additional contributions compared to Taamneh's dataset. We introduce more sensors such as multiple RGB, thermal and infrared cameras of variant qualities, capturing different points of view, as well as additional physiological markers such as the raw blood volume pulse, respiration and skin temperature measurements. In

addition, we explore some more realistic scenarios such as the GPS interaction and the radio listening events, which are some of the most common activities that occur in today's driving. Lastly, the dataset here is designed to promote parallel investigation of drowsy and distracted driving states; a characteristic that makes this resource quite unique. Even though drowsiness characterization is out of the scope of this paper the data and the available labels have been specifically designed to support such research in the future.

Most recently in early 2020, the paper by McDonald et al. [25] discussed the advantage of ensemble learners to model driver behavior and classify distractors based on physiological markers. Overall though, physiological data has 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 [7, 53].

Thermal imaging was also investigated by researchers as a non-invasive modality. Avinash et al. [3] presented a research that sought to monitor the driver's distractions using a ThermoVision SC6000 **Mid-Wave Infrared (MWIR)** camera. The authors suggested an approach that is based on the face's thermal signature of 11 individuals. They performed two experiments to evaluate the validity of their method. The first experiment aimed to model the driver's cognitive distraction by permitting mobile phone usage while driving and the second experiment focused on the visual distraction of drivers by permitting texting while driving. In order to extract the facial signature, a smoothie tracker was applied to track the supraorbital region of the subjects. Their work showed the potential of thermal imaging based on the facial physiological monitoring system in detecting driver's distraction. The works proposed by Kolli et al. [17] in 2011 used an infrared thermal camera ("PathFindIR") from FLIR systems with spectral band 8-14μwith the aim of classifying driver's emotions. In addition to detecting driver's anger, disgust, fear, joy, sadness, and surprise, the authors developed three different algorithms after analysing hundreds of thermograms to identify the face region. In 2017, in order to classify individuals as fatigued or resting, Lopez et al. [23] processed thermal images following three primary steps using a Therm-App mobile thermal camera. In the first step, three sub steps were included which are detection, segmentation, and alignment of thermal facial regions by using the position of the eyes and nose. The alignment of the images aimed to reduce any potential discrepancy between the subjects and images. This sub step produced a collection of aligned thermal facial images as well as regions of interest. The second step employed two separate convolutional neural networks to generate fixed-length deep feature vectors extracted from facial images and regions. The third step then utilized these features with a **Support Vector Machine (SVM)** to determine if a subject is fatigued or resting. In 2019, the research done by Knapik et al. [16] proposed a unique method for detecting yawns using long-range infrared imaging. Their results showed a great potential in detecting driver's fatigue in both laboratory and real car conditions. More recently in 2021, Schif et al. [42] measured the local temperature variation distribution in order to detect sweating by characterizing surface roughness. This method could be used to avoid driver's distraction by regulating the car's climate. Researchers also exploited the potential of thermal imaging in detecting the driver's physical state. For instance, Forczmanski et al. [11] showed that thermal image analysis could be exploited to estimate the state of the eyes and mouth, which can be used to detect driver's drowsiness.

While several recent papers have tried to study the fluctuations of stress during driving [51, 52], very few have focused explicitly on how different common driving distractors affect specific physiological and behavioral reactions [55], and even fewer have explored the potential of multimodal data for such purposes [6, 37].

This paper tries to fill some of the gaps that past research has not targeted extensively yet. Firstly, we explore the problem of distraction detection. Next, we aim to understand how different

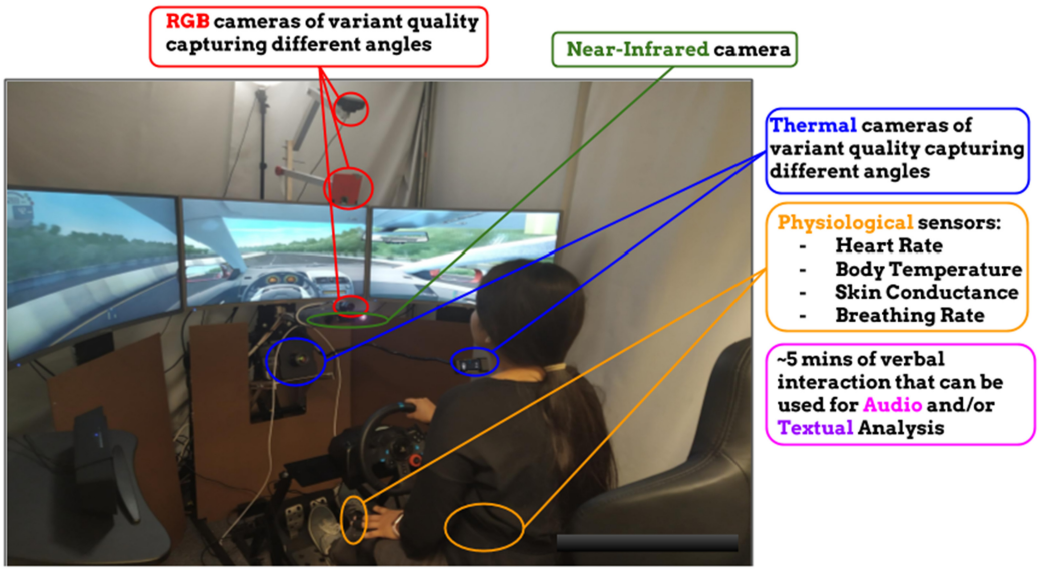


Fig. 1. The data collection experimental setup.

distractions can be discriminated by addressing the problem of distraction recognition. Our evaluation goes beyond physical distractions and tries 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 hope to identify features that can be evaluated in the future to analyse distractions and their impact on human performance in applications other than driving such as in education, training and other task-oriented domains [33]. Secondly, we aim to explore the advantages of three different modalities towards identifying different distractions. Our results highlight the advantage of our approaches on distraction recognition and offer 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 routines so that they are as least 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 psychology and cognitive neuroscience, designed to measure working memory [15]. 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 prerecorded 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. 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 [4]. 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 [41]. 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 performed experiments 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 [50].

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 7.

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. A smaller window size was selected as per the findings from Lee et al. [19], which found that early warning times greatly reduced collisions caused by driver distraction. For every window we extract the following features describing the distribution of AU intensities within a window; minimum and maximum values, average, variance, skewness and kurtosis. 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 [26]. 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 breaths 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, i.e., 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 eight 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. Past research also found that physiological signals provided better feature quality and performance with smaller window sizes [36, 54]. At every temporal window we compute for every feature the same six statistics mentioned in Section 4.1 plus the zero-crossing rate [38]. 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 1,078 features representing four seconds of physiological measurements.

4.3 Thermal Data Processing

To analyze the thermal features, we first located five different regions, including the whole face, forehead, eyes, cheeks, and nose. Afterwards, these regions were tracked throughout the thermal videos by applying the tracking algorithm proposed in [46]. Specifically, the process is divided into three steps: face segmentation, tracking, and the creation of a thermal map, using our approach provided in [1]. The final step consists of extracting statistical features for all **Regions Of Interest (ROIs)** to form a thermal map in addition to segmenting the thermal features using three window sizes of two, four and eight seconds, each with a 50% overlap.

Looking more closely at these steps, we began by manually locating the ROIs by defining their bounding boxes in the first frame, as automatic facial detection methods, including contour tracking methods and template matching, did not perform well on thermal images. Thereafter, points of interest in the detected ROIs were captured using a variation of the Shi-Tomasi corner detection algorithm [44] by computing the weighted square difference between two successive frames. As the method compares an image patch $I_1(x_i)$ with a shifted version of the image, $I_1(x_i + \Delta u)$, an auto-correlation function S was used.

$$S(\Delta u) = \sum_i w(x_i)(I_0(x_i + \Delta u) - I_0(x_i))^2 \tag{1}$$

where u is the displacement vector and $w(x_i)$ is a window function. The function is approximated using Taylor Series expansion into

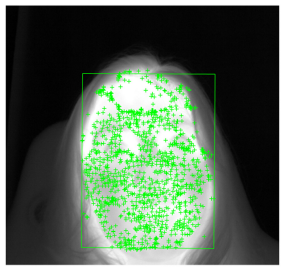


Fig. 2. Points of interest detected in the face region.

$$S(\Delta u) \approx \sum_i w(x_i)(\nabla I_0(x_i).\Delta u)^2 \tag{2}$$

where,

$$\nabla I_0(x_i) = \left(\frac{\partial I_0}{\partial x}, \frac{\partial I_0}{\partial y}\right)(x_i) \tag{3}$$

We used a fixed-size Gaussian filter to smooth the calculated gradient. Thus, S can be rewritten as:

$$S(\Delta u) = \Delta u^T V \Delta u \tag{4}$$

where V denotes the auto-correlation matrix. The interesting corner points to be tracked were located using the variation in S by computing the minimum eigenvalues from V. Figure 2 highlights the points of interest discovered in the facial region with a lower threshold, allowing for more detected points. These points suggested the presence of a blood vein regulating the temperature of the surrounding region where sharper changes in the colors were present.

We tracked the ROI bounding box for the duration of the videos using a fast version of the the **Kanade-Lucas-Tomasi (KLT)** tracking method [46], which provides accurate results for stabilizing the ROI bouding box. The tracking algorithm estimates the relocation of points of interest between two successive frames by assuming a small displacement between pixels in a frame at time t and t + τ, which was ideal for our tracking needs. Afterwards, a geometric transformation was applied to estimate the transformation of interesting points based on similarity between the frames. As a precaution, we established a threshold of 95% of correctly mapped points between two successive frames to account for any occlusion, such as not having the subject face in the frame or just obtaining it partially. In the occurrence of occlusion, the current frame is skipped and tracking resumes to the following frame. Lastly, for each ROI, we generated a thermal map that reflected the thermal distribution. This was accomplished through the steps of ROI segmentation, Segment binarization, Image masking, and finally Thermal map cropping. This process is illustrated in Figure 3.

For each of the five ROIs, we thereby extract a total of 24 features, including 20 histogram related and four statistical measures derived from those regions. The histogram features describe the temperature distribution in the ROI over 20 bins, while the four statistical features represent the mean temperature, the range of temperatures, the minimum temperature value, and the maximum temperature value, all taken per frame. Next, we segment the thermal features using three window sizes of two, four and eight seconds respectively, with a a 50% overlap for both, which were selected in order to be compared against the RGB and Physiological modalities. For each window, six statistical features were collected as desrcibed in Sections 4.1 and 4.2, those being the mean, minimum and maximum, variance, skewness and kurtosis. We end the extraction process with a feature vector for each window comprising of $5 \times 24 \times 6 = 720$ features in total. Using

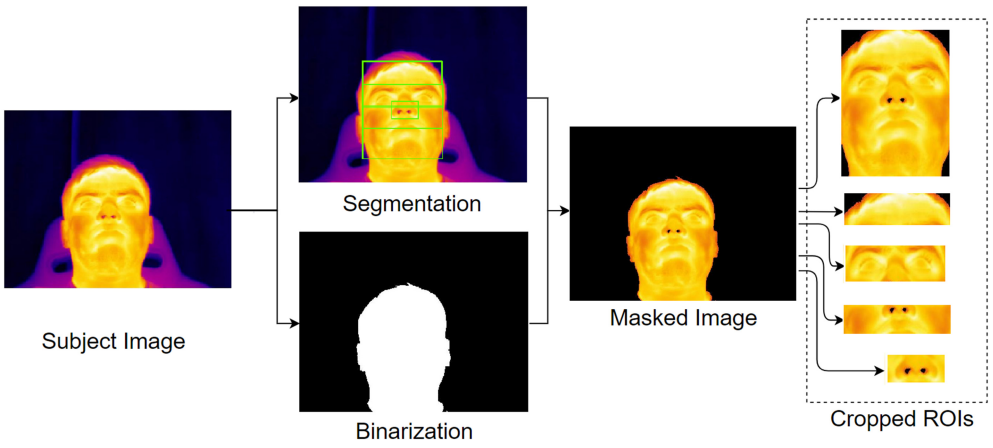


Fig. 3. Segmenting, binarizing, masking and cropping the thermal ROI.

the full dataset of recordings, we arrive at approximately 17 hours of free-driving based thermal data and approximately 9 hours of distractor based thermal data at our disposal for training and classification.

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 used entropy as a metric to ensure maximum information gain at each node [21].
(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 used 100 weak estimators [12].

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 [25]. 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, D3-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. After the quantitative analysis provided by the classification results we discuss a qualitative evaluation that aims to identify features that contributed the most.

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

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

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**

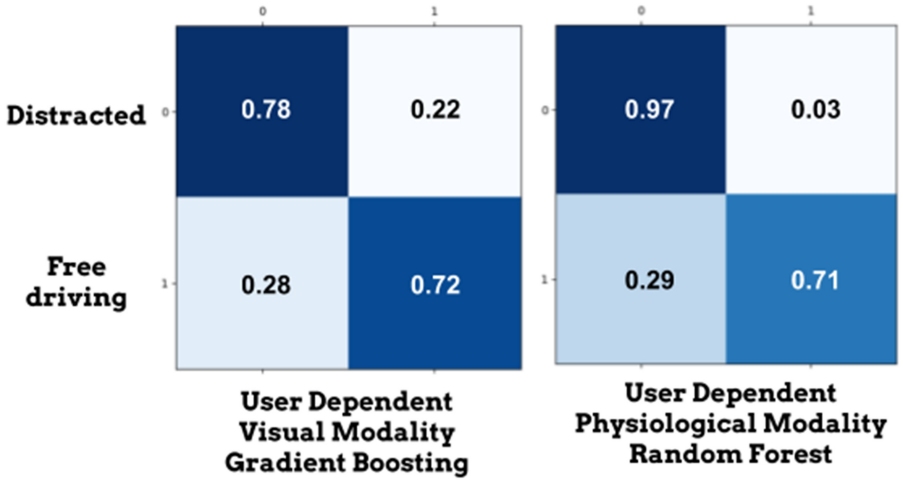


Fig. 4. Confusion matrices on distracted VS non-distracted driving classification for the best results of Table 2.

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 4 illustrates the confusion matrices for the best results using each modality.

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 4 reveal that the physiological model greatly 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. Table 3 shows the results on each experiment, and Figures 5 and 6 show their corresponding confusion matrices.

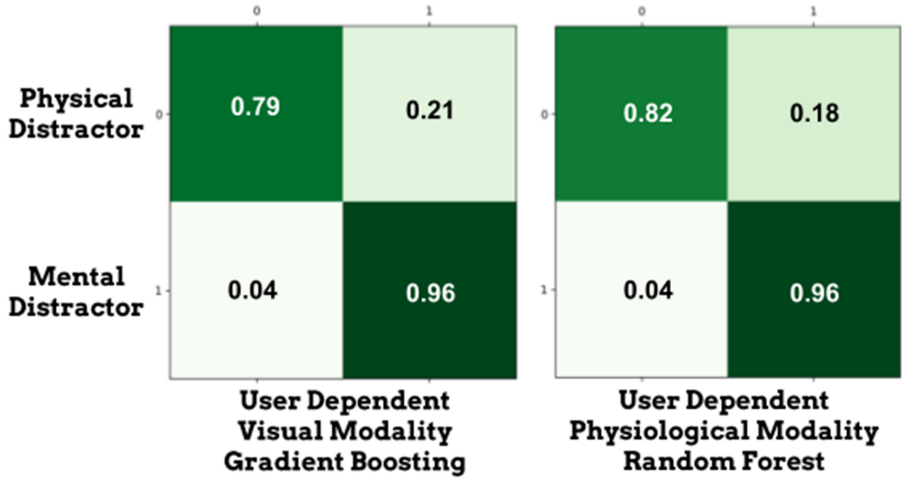


Fig. 5. Confusion matrices on distraction recognition as a 2-class problem for the best results of Table 3.

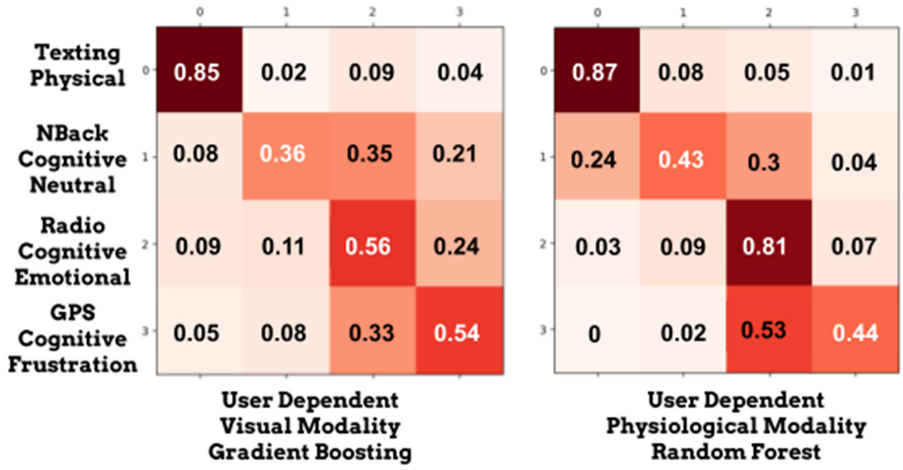


Fig. 6. Confusion matrices on distraction recognition as a 4-class problem for the best results of Table 4.

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**

Overall, all unimodal modeling approaches in Table 3 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,

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 of Distraction Recognition as a 4-class Problem with Respect to Precision

	Visual		Physiological	
	RF	GB	RF	GB
User Independent	0.54	0.54	0.67	0.4
User Dependent	0.58	0.59	0.69	0.65
User Exclusive	0.51	0.47	0.33	0.32

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 considerable behavioral differences across subjects, which translates to the minor increase in performance in the 'User Independent' and 'User 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 appreciable 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 6. We can also see from Table 5 the improved precision when using the physiological sensors for distraction recognition.

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

Figure 7 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 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 [34]. 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 8(a) and 8(b) show feature importance values for each feature, i.e., 102 visual and 1,078 physiological features (see Sections 4.1, 4.2).

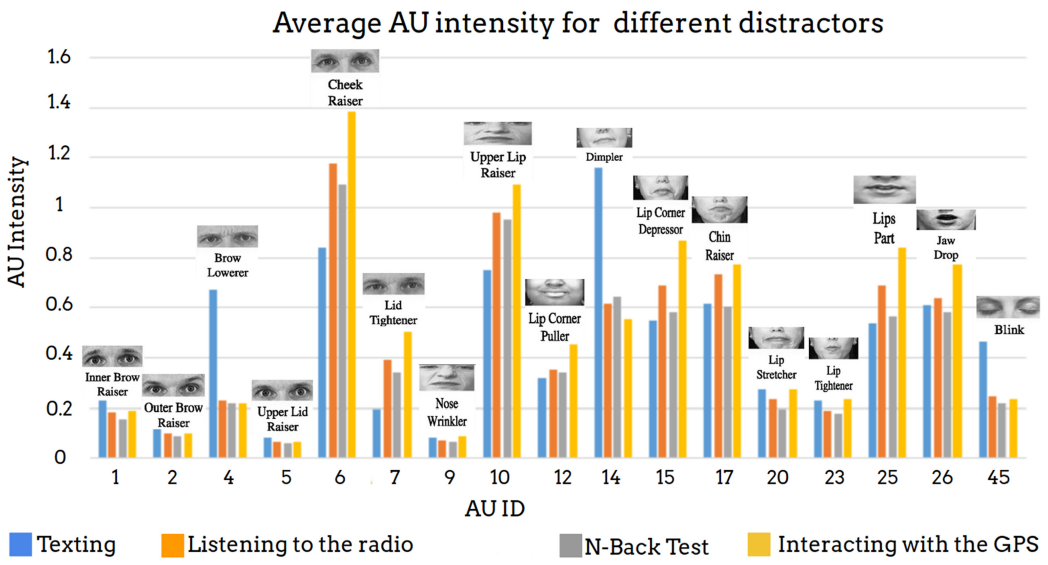


Fig. 7. AU Intensities per distractor. AU ID numbers are defined by FACS.

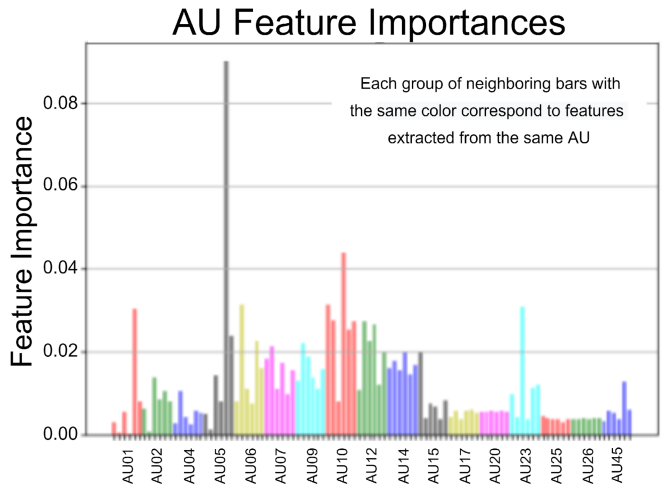
Table 6. Top Five Features of Each Modality Based on Feature Importance

	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

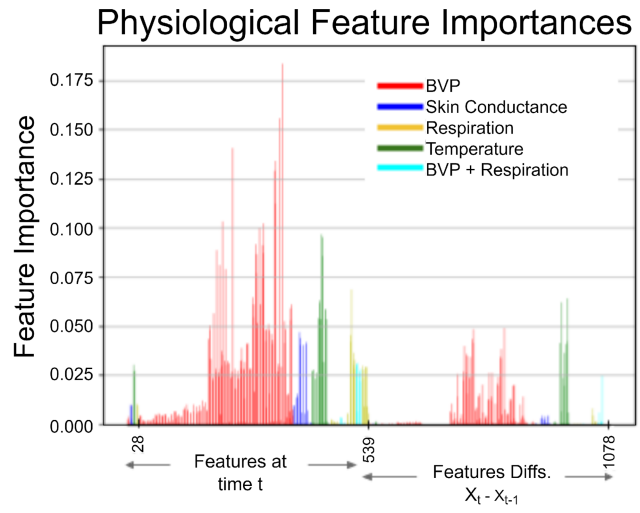
Table 6 presents the top#5 performing "core features" from each signal. By "core features", we refer to the initial intensities for the 17 AUs and the 77 domain-specific physiological features before extracting window-based statistics.

It is worth observing that across the visual features (Figure 8(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 8(b)), BVP related features seem to account the most for the good results offered by the modality. However, as seen in Figure 8(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 below:

- 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.



(a) Feature importance for Visual Features



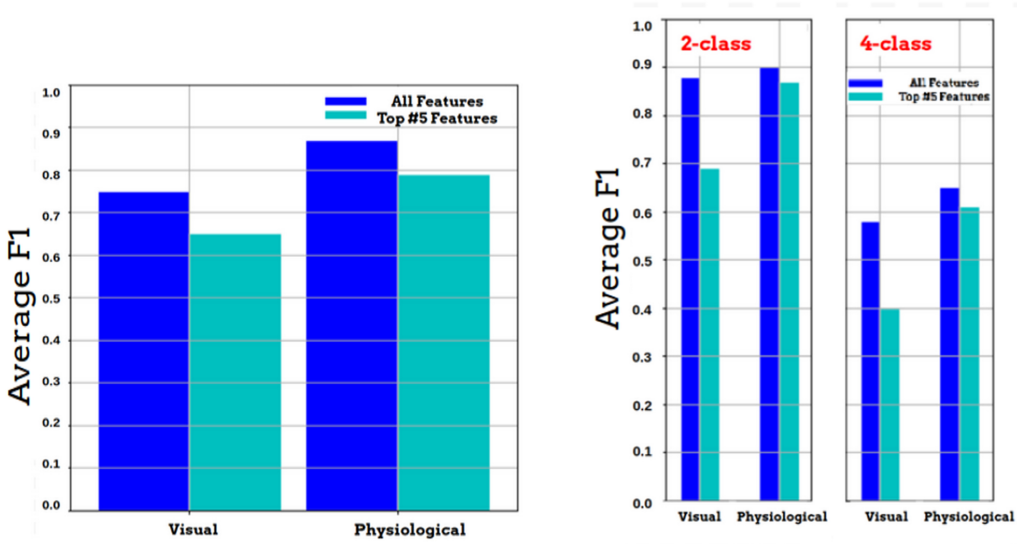
(b) Feature importance for Physiological Features

Fig. 8. Feature importance for each modality based on information gain.

pNN features are known to be highly correlated with sympathetic and parasympathetic modulation of the nervous system [13, 39]. 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 [20].

To emphasize the importance 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 9(a) and 9(b).

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



(a) Distraction Detection (b) Distraction Recognition

Fig. 9. Performance comparison between best models (User-Dependent scheme) trained on the top five features of each modality VS all the available features.

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

5.3 Performance of the Thermal Modality in Capturing Distracted Behavior

For the thermal modality we use the same experimental setups as discussed above for the visual and physiological modalities, carrying out three experiments:

(1) Binary distraction detection between free-driving and distracted driving.
(2) 2-class distraction recognition between physical and mental distractors.
(3) 4-class distraction recognition between physical, cognitive, emotional and frustration distractors.

For each experiment we evaluate the performance using the average F1-score against two baselines, one single label and the balanced label, as discussed earlier in Section 5.1. Performance across three window sizes are observed, all using a 50% overlap between segments. Similar to the experiments for the visual and physiological modalities, three modeling approaches, including User Independent, User Dependent and User Exclusive are used. Leave one subject out cross validation is used to assess the performance.

Table 7 outlines the results obtained when using each of the three modeling approaches over the three window sizes for the distraction recognition problem. We can observe that a longer 8-second window is more beneficial to User Independent and Dependent modeling approaches. However, User Exclusive benefits from smaller 2-second window sizes. This implies that highly personalized detection might be more likely to be classified more efficiently.

RF classifiers are the best performing classifiers for all modelling approaches, with the User Dependent modelling with an 8-second window achieving the best performance with an average

Table 7. Results on Distracted VS Non-distracted Driving Classification with Respect to Average F1 using the Thermal Modality

	Baseline		Thermal					
			2 second window, 50% overlap		4 second window, 50% overlap		8 second window, 50% overlap	
	Single Label	Balanced	RF	GB	RF	GB	RF	GB
User Independent	0.32	0.5	0.704	0.704	0.71	0.715	0.743	0.735
User Dependent	0.32	0.5	0.711	0.716	0.738	0.73	0.747	0.746
User Exclusive	0.32	0.5	0.619	0.608	0.617	0.606	0.616	0.578

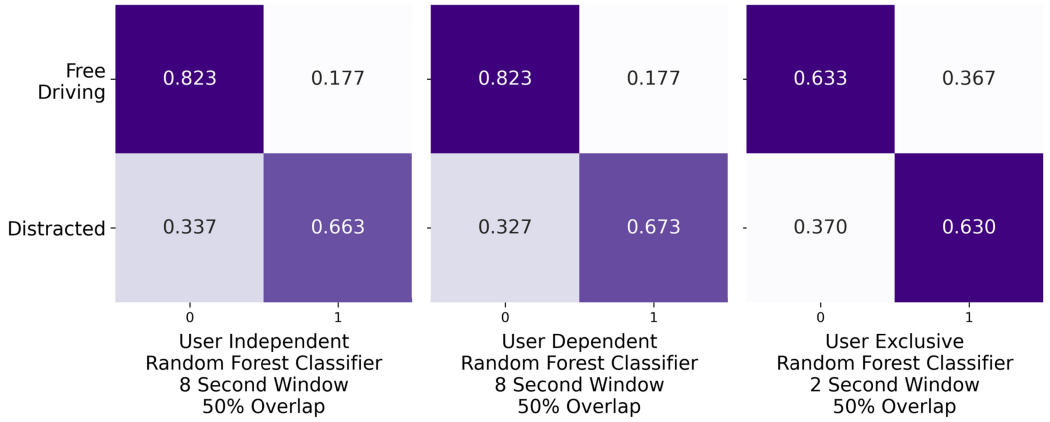


Fig. 10. Confusion matrices on distracted VS non-distracted driving classification for the best results of Table 7 for each modeling approach.

F1-score of 74.71%. Looking at the confusion matrices in Figure 10, we observe that the thermal modality is much better at classifying freedriving over distraction. This, however, could be an effect of a class imbalance, as freedriving recordings are at a 1.8:1 ration to the distracted recordings. In this context, it is possible having more distractor data to balance the classes would improve the performance.

As seen in Table 8, we observe that the thermal modality performs the best in 2-class distraction recognition, with an average F1-score as high as 94.21% when using an 8-second window in a User Dependent modelling approach. A similar trend of benefiting the User Exclusive approach using smaller window sizes is seen here as well, where the GB classifier achieves an F1-score of 86.29% when using 2-second windows. All modeling approaches have better classification for the mental distractor class, as seen in the confusion matrices in Figure 11. It should be noted that for User Independent and Dependent modeling, the imbalance is much lower compared to the distraction detection experiment.

Finally, for the third experiment, a 4-class distraction recognition problem, the results are tabulated in Table 9. Here, we see that the GB classifier is the best performer for all modelling approaches, with the User Dependent 8-second window approach achieving an F1-score of 79.68%. By looking at the confusion matrices in Figure 12 we see that the models are the best at correctly identifying the physical and frustration distractors. This could be happening due to an increase in the subjects' movements and more exclusive variations in expressions that are captured in these two distractors over the cognitive and emotional ones.

Table 8. Results on Distraction Recognition as a 2-class Problem with Respect to Average F1 using the Thermal Modality

	Baseline		Thermal					
			2 second window, 50% overlap		4 second window, 50% overlap		8 second window, 50% overlap	
	Single Label	Balanced	RF	GB	RF	GB	RF	GB
User Independent	0.32	0.5	0.933	0.928	0.935	0.932	0.94	0.936
User Dependent	0.32	0.5	0.934	0.929	0.938	0.935	0.942	0.94
User Exclusive	0.32	0.5	0.699	0.863	0.69	0.798	0.688	0.708

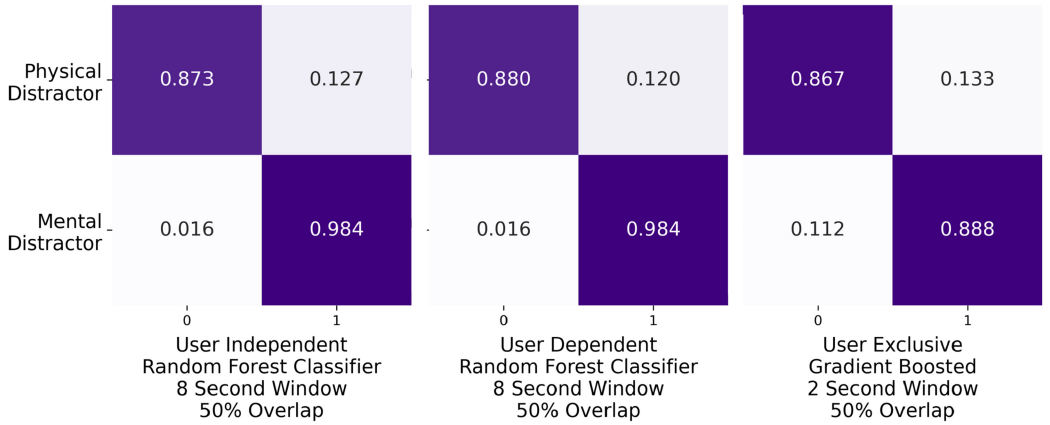


Fig. 11. Confusion matrices on distraction recognition as a 2-class problem for the best results of Table 8 for each modeling approach.

At this stage, we observe some key trends consistent across all experiments:

(1) User Independent and Dependent modeling approaches are by far the stronger modelling approaches compared to the User Exclusive one. This is possibly due to the much larger amount of training data available for these approaches, which aids in developing much more robust and generalized models. On the other hand, the User Exclusive modelling approach gets to use only a single subject's data for training, severely limiting the scope of data available for use.
(2) User Dependent modeling is slightly better in performance compared to User Independent modeling in all experiments. This is most likely due to the model gaining specific intuitive knowledge for a given subject when it is allowed to use some of the subject's data during training. However, the increase in performance is not too great, in the range of 0.3-0.5%.
(3) While the longer 8-second window generates the best results, the 4-second window is only 2% short of matching those results. The small trade off in performance to halve the classification time can prove valuable for scenarios where a real-time prediction is required on the expense of a small loss in accuracy.

5.4 Performance of ROIs in the Thermal Modality for Classification

Looking at the performance of the thermal modality in detecting and recognizing distractions, Figure 13 outlines the features that were selected to be the most important for classification for

Table 9. Results on Distraction Recognition as a 4-class Problem with Respect to Average F1 using the Thermal Modality

	Baseline		Thermal					
			2 second window, 50% overlap		4 second window, 50% overlap		8 second window, 50% overlap	
	Single Label	Balanced	RF	GB	RF	GB	RF	GB
User Independent	0.1	0.25	0.728	0.739	0.744	0.757	0.758	0.784
User Dependent	0.1	0.25	0.746	0.76	0.762	0.773	0.768	0.797
User Exclusive	0.1	0.25	0.442	0.561	0.424	0.492	0.413	0.422

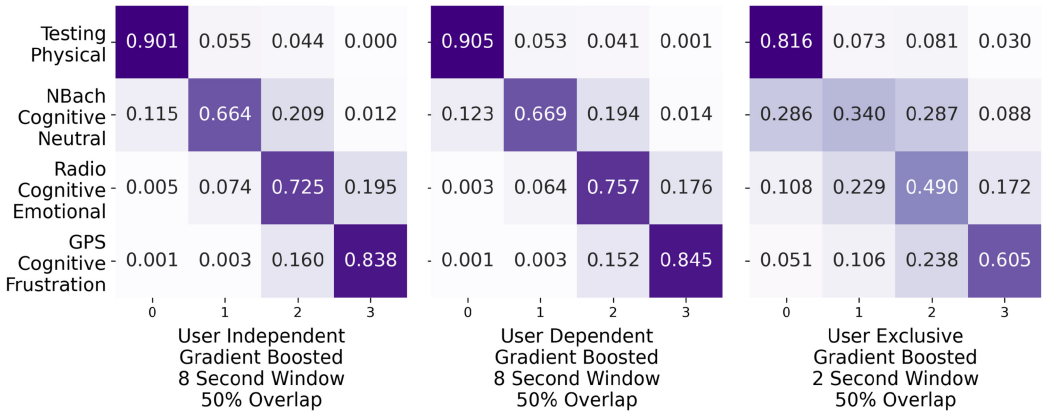


Fig. 12. Confusion matrices on distraction recognition as a 4-class problem for the best results of Table 9 for each modeling approach.

each experiment respectively. Lasso Leave One Subject Out Cross Validation was used with the RF classifier to determine feature selection.

For both the 2-class and 4-class distraction recognition, we see that the cheeks are the most important ROI contributing to the majority of the features used. The cheeks being a strong performer likely indicates that thermal imaging, being based on temperature readings, would benefit from clear visibility of the skin for accurate measurement. However, for distraction recognition, the cheeks are not the single most important ROI, with the face and nose being contributors to the selected features as well.

However, taking a look at the ROI performance when using only one ROI at a time instead of the fusion normally used for the modality as a whole in Figures 14 and 15, respectively, we see that the individual ROI performances do not correlate with the feature importance. In almost all experiments regardless of window size, it was the face that performed the best, especially in distraction detection and 4-class distraction recognition.

The eyes are the best performing ROI in the 2-class distraction recognition, however, implying that this ROI is suitable for differentiating between physical and mental distractors once distraction has been detected. The eyes region include ducts that carry a large amount of blood indicating the possibility that the blood flow in this region changes between physical and mental distractions, resulting in different temperature patterns. The cheeks were in no case the best ROI for detection or recognition, coming in second or third for performance.

In all cases, using the thermal modality with all ROIs together results in equal or better performance for all experiments and modelling approaches when compared to using individual ROIs.

Fig. 13. Feature importance for Thermal modality based on Lasso CV.

Table 10. Comparison of Results between the Three Modalities When Using the User Dependent Modelling Approach

	Visual	Physiological	Thermal
Distraction Detection	0.75	0.87	0.747
2 Class	Gradient Boosted	Random Forest	Random Forest
Distraction Recognition	0.88	0.9	0.94
2 Class	Gradient Boosted	Random Forest	Random Forest
Distraction Recognition	0.58	0.65	0.797
4 Class	Gradient Boosted	Random Forest	Gradient Boosted

While certain ROIs might have a better affinity towards certain kinds of detection or recognition tasks, their combination as a whole is still much more suited as a universal approach for distraction detection and recognition.

5.5 Comparison of the three Modalities in Distraction Detection and Recognition Tasks

Table 10 highlights the best F1-score obtained per modality. In all cases the User Dependent approach is the best modelling approach regardless of aiming for detection or recognition. We can observe that for distraction detection, the physiological modality is the best performer with an

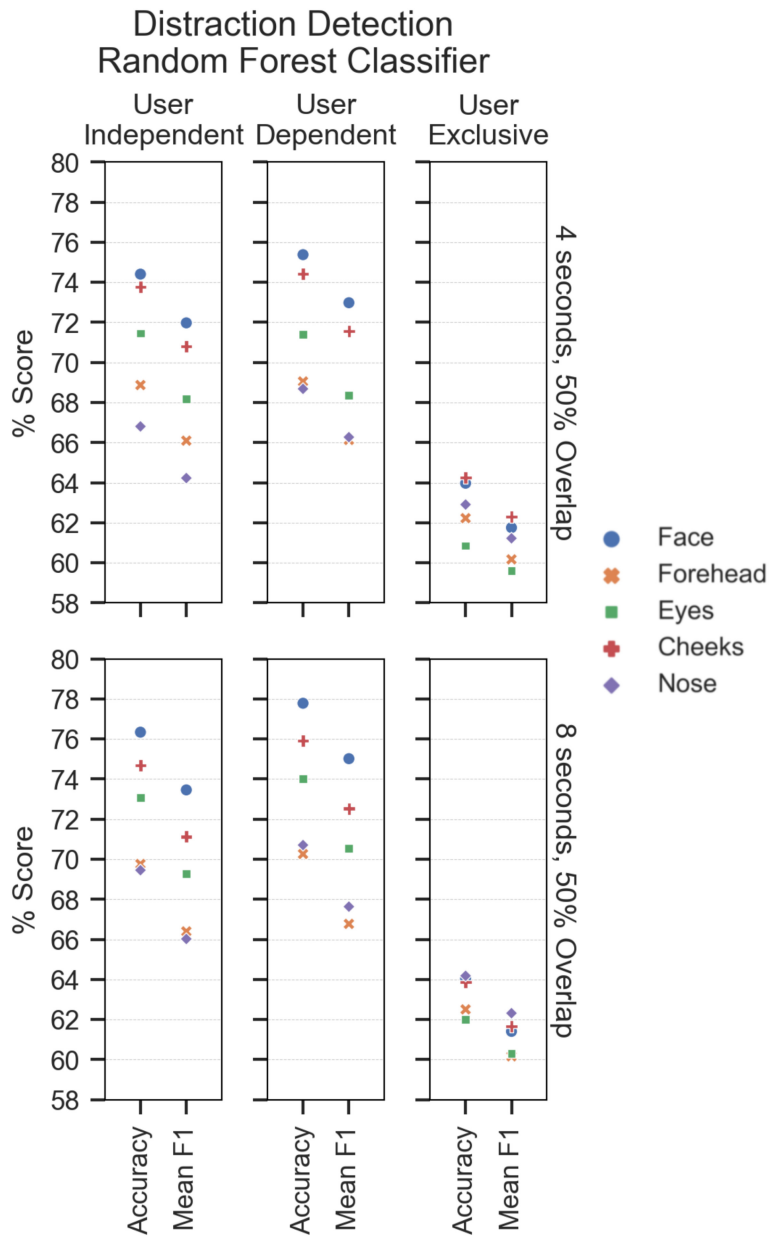


Fig. 14. Performance metrics on distracted VS non-distracted driving classification for each ROI in the Thermal modality.

87% F1-score. However, in case of distraction recognition, the thermal modality outperforms the visual and physiological modalities, with F1-scores of 94.7% and 79.7% in the 2-class and 4-class recognition tasks, respectively. The thermal modality outperforms other modalities for the 4-class distraction recognition in all respects, being able to identify individual distractors more accurately than the other modalities. Additionally, the physiological modality has a better recall for the distraction class in distraction detection whereas the thermal modality achieves a higher recall when

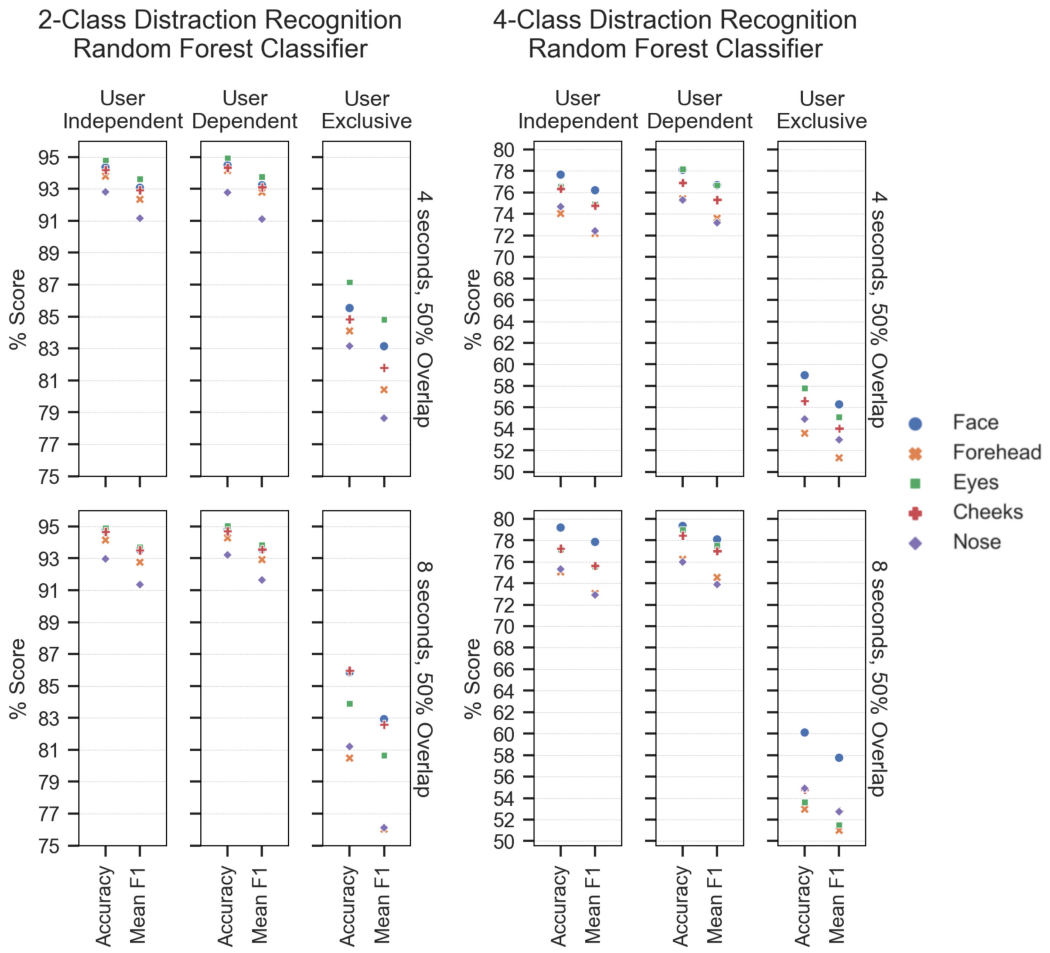


Fig. 15. Performance metrics on distraction recognition for each ROI in the Thermal modality for each modeling approach.

it comes to the freedriving class. This indicates that the physiological signals show greater performance in case of the presence of any type of drivers' distraction. However, the thermal modality is expected to show further improvement in detecting distraction with more balanced instances among the two classes. The thermal maps provide better discrimination between the different types of distractors once distraction is detected. This also indicates the potential of using thermal imaging as a non-contact and less invasive approach to detect and recognize distraction.

Table 11 details some of the F1 sores that we obtained when performing preliminary experiments on the data using a SVM with a RBF kernel. As discussed in 4.4, we chose to not use this classifier as its performance is almost always superseded by those using ensemble classifiers alongside a much longer training time.

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

Table 11. Results when using the SVM-RBF Classifier Across the Visual (V), Physiological (P), and Thermal 8-second Window (T) Modalities with Respect to Average F1

	Distraction Detection, 2-class			Distraction Recognition, 2-class			Distraction Recognition, 4-class		
	V	P	T	V	P	T	V	P	T
User Independent	0.68	0.85	0.75	0.85	0.86	0.91	0.52	0.64	0.75
User Dependent	0.68	0.85	0.77	0.86	0.87	0.91	0.53	0.65	0.76
User Exclusive	0.65	0.52	0.65	0.81	0.31	0.37	0.48	0.17	0.42

the experimental nature of our setup, there is substantial research evidence to support the direct application and integration of our methods in modern vehicles [8, 14, 32, 43, 47].

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 contributions 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 physiological 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.

How does the thermal modality perform in detecting and recognizing distracted behavior? The thermal modality is more attuned towards detecting driver attentiveness more accurately than it is with distraction. This could be an effect of the class imbalance. Within the task of distraction recognition as a 2-class problem, it performs very well in distinguishing between physical and mental distractors. When modelling distraction recognition as a 4-class problem, the modality performs better in detecting physical and frustration distractors, falling behind on the cognitive and emotional ones. The primary ROIs contributing to classification are the face, eyes and the cheek, but the fusion of all five ROIs provide better performance metrics for all the conducted experiments.

How does each modality compare for the distraction detection and recognition problems? We found that the physiological modality is suited for distraction detection, with a better recall in detecting distraction. However, with a more balanced data in terms of classes, we expect the thermal modality to show further improvement. For distraction recognition, the thermal modality is the best performing modality for both 2-class and 4-class tasks, performing with up to 94% F1-score in the 2-class variant.

In conclusion, we presented a novel dataset consisting of thermal, visual and physiological modalities on which we performed experiments to understand the applications of distracted behavior detection using machine learning. More specifically, we found that the visual and physiological modalities are better at distraction detection, with the physiological modality reaching F1-scores of 81% in this task. Finally, we looked at the thermal modality, observing that it was the best modality for distraction recognition, with 94% F1-scores in the 2-class variant of the task.

REFERENCES

[1] Mohamed Abouelenien, Verónica Pérez-Rosas, Rada Mihalcea, and Mihai Burzo. 2017. Detecting deceptive behavior via integration of discriminative features from multiple modalities. *IEEE Transactions on Information Forensics and Security* 12, 5 (2017), 1042–1055.

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

[3] W. Avinash, S. Dvijesh, and P. Ioannis. 2010. A novel method to monitor driver's distractions. In *Proceedings of the 28th International Conference Extended Abstracts on Human Factors in Computing Systems, Atlanta, Georgia, USA*.

[4] 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.

[5] Shahina Begum. 2013. Intelligent driver monitoring systems based on physiological sensor signals: A review. In *16th International IEEE Conference on Intelligent Transportation Systems (ITSC'13)*. IEEE, 282–289.

[6] 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

[7] Lan-lan Chen, Yu Zhao, Jian Zhang, and Jun-zhong Zou. 2015. Automatic detection of alertness/drowsiness from physiological signals using wavelet-based nonlinear features and machine learning. *Expert Systems with Applications* 42, 21 (2015), 7344–7355.

[8] 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.

[9] Chelsea Dobbins and Stephen Fairclough. 2018. Detecting negative emotions during real-life driving via dynamically labelled physiological data. In *2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. IEEE, 830–835.

[10] 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.

[11] Paweł Forczmański and Anton Smoliński. 2021. Supporting driver physical state estimation by means of thermal image processing. In *International Conference on Computational Science*. Springer, 149–163.

[12] Jerome H. Friedman. 2002. Stochastic gradient boosting. *Computational Statistics & Data Analysis* 38, 4 (2002), 367–378.

[13] 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.

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

[15] Michael J. Kane, Andrew R. A. Conway, Timothy K. Miura, and Gregory J. H. 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.

[16] Mateusz Knapik and Bogusław Cyganek. 2019. Driver's fatigue recognition based on yawn detection in thermal images. *Neurocomputing* 338 (2019), 274–292.

[17] Abhiram Kolli, Alireza Fasih, Fadi Al Machot, and Kyandoghere Kyamakya. 2011. Non-intrusive car driver's emotion recognition using thermal camera. In *Proceedings of the Joint INDS'11 & ISTET'11*. IEEE, 1–5.

[18] 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.

[19] John D. Lee, Daniel V. McGehee, Timothy L. Brown, and Michelle L. Reyes. 2002. Collision warning timing, driver distraction, and driver response to imminent rear-end collisions in a high-fidelity driving simulator. *Human Factors* 44, 2 (2002), 314–334. https://doi.org/10.1518/0018720024497844 arXiv:https://doi.org/10.1518/0018720024497844 PMID: 12452276.

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

[21] Andy Liaw, Matthew Wiener, et al. 2002. Classification and regression by Random Forest. *R News* 2, 3 (2002), 18–22.

[22] 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.

[23] Miguel Bordallo Lopez, Carlos R del Blanco, and Narciso Garcia. 2017. Detecting exercise-induced fatigue using thermal imaging and deep learning. In *2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*. IEEE, 1–6.

[24] 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.

[25] 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.

[26] H. Meyers. 2010. ProComp infiniti/biograph infiniti biofeedback system (version 5.1. 2). *Montreal, QB: Thought Technology Ltd* (2010).

[27] Stephan Mühlbacher-Karrer, Ahmad Haj Mosa, Lisa-Marie Faller, Mouhannad Ali, Raiyan Hamid, Hubert Zangl, and Kyandoghere Kyamakya. 2017. A driver state detection system-combining a capacitive hand detection sensor with physiological sensors. *IEEE Transactions on Instrumentation and Measurement* 66, 4 (2017), 624–636.

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

[29] 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. (2019). [Online; accessed 13-April-2020].

[30] 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. (2020). [Online; accessed 13-April-2020].

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

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

[33] 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.

[34] 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.

[35] 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.

[36] Mohammad Naim Rastgoo, Bahareh Nakisa, Frederic Maire, Andry Rakotonirainy, and Vinod Chandran. 2019. Automatic driver stress level classification using multimodal deep learning. *Expert Systems with Applications* 138 (2019), 112793. https://doi.org/10.1016/j.eswa.2019.07.010

[37] 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 (PETRA'20)*. Association for Computing Machinery, New York, NY, USA, Article 12, 9 pages. https://doi.org/10.1145/3389189.3389192

[38] 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.

[39] 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.

[40] Wang Rongben, Guo Lie, Tong Bingliang, and Jin Lisheng. 2004. Monitoring mouth movement for driver fatigue or distraction with one camera. In *Proceedings of the 7th International IEEE Conference on Intelligent Transportation Systems (IEEE Cat. No. 04TH8749)*. IEEE, 314–319.

[41] 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.

[42] Diana Schif, Holger Forst, and Ulrich T. Schwarz. 2021. Methods for sweat detection in cars based on thermal images of the face. *IEEE Sensors Journal* (2021).

[43] 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.

[44] Jianbo Shi et al. 1994. Good features to track. In *1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 593–600.

[45] Heung-Sub Shin, Sang-Joong Jung, Jong-Jin Kim, and Wan-Young Chung. 2010. Real time car driver's condition monitoring system. In *Sensors, 2010 IEEE*, 951–954.

[46] Sudipta N. Sinha, Jan-Michael Frahm, Marc Pollefeys, and Yakup Genc. 2006. GPU-based video feature tracking and matching. In *EDGE, Workshop on Edge Computing Using New Commodity Architectures*, Vol. 278. 4321.

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

[48] 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.

[49] Salah Taamneh, Panagiotis Tsiamyrtzis, Malcolm Dcosta, Pradeep Buddharaju, Ashik Khatri, Michael Manser, Thomas Ferris, Robert Wunderlich, and Ioannis Pavlidis. 2017. A multimodal dataset for various forms of distracted driving. *Scientific Data* 4 (2017), 170110.

[50] 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.

[51] 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'19)*, 1–10.

[52] 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.

[53] 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).

[54] Shuo Yang, Zhong Yin, Yagang Wang, Wei Zhang, Yongxiong Wang, and Jianhua Zhang. 2019. Assessing cognitive mental workload via EEG signals and an ensemble deep learning classifier based on denoising autoencoders. *Computers in Biology and Medicine* 109 (2019), 159–170. https://doi.org/10.1016/j.compbiomed.2019.04.034

[55] 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

Received 9 August 2021; revised 16 November 2021; accepted 15 February 2022