# The Empathetic Car: Exploring Emotion Inference via Driver Behaviour and Traffic Context

SHU LIU*, ETH Zürich, Switzerland
KEVIN KOCH, University of St. Gallen, Switzerland
ZIMU ZHOU, Singapore Management University, Singapore
SIMON FÖLL, ETH Zürich, Switzerland
XIAOXI HE, ETH Zürich, Switzerland
TINA MENKE, Karlsruhe Institute of Technology, Germany
ELGAR FLEISCH, ETH Zürich, Switzerland
FELIX WORTMANN, University of St. Gallen, Switzerland

An empathetic car that is capable of reading the driver's emotions has been envisioned by many car manufacturers. Emotion inference enables in-vehicle applications to improve driver comfort, well-being, and safety. Available emotion inference approaches use physiological, facial, and speech-related data to infer emotions during driving trips. However, existing solutions have two major limitations: Relying on sensors that are not built into the vehicle restricts emotion inference to those people leveraging corresponding devices (*e.g.*, smartwatches). Relying on modalities such as facial expressions and speech raises privacy concerns. By contrast, researchers in mobile health have been able to infer affective states (*e.g.*, emotions) based on behavioral and contextual patterns decoded in available sensor streams, *e.g.*, obtained by smartphones. We transfer this rationale to an in-vehicle setting by analyzing the feasibility of inferring driver emotions by passively interpreting the data streams of the control area network (CAN-bus) and the traffic context (inferred from the front-view camera). Therefore, our approach does not rely on particularly privacy-sensitive data streams such as the driver facial video or driver speech, but is built based on existing CAN-bus data and traffic information, which is available in current high-end or future vehicles. To assess our approach, we conducted a four-month field study on public roads covering a variety of uncontrolled daily driving activities. Hence, our results were generated beyond the confines of a laboratory environment. Ultimately, our proposed approach can accurately recognise drivers' emotions and achieve comparable performance as the medical-grade physiological sensor-based state-of-the-art baseline method.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing**.

Additional Key Words and Phrases: Emotion recognition, Driving behaviours, Traffic contexts, Control area network (CAN), Intelligent vehicle

*Corresponding author

Authors' addresses: Shu Liu, liush@ethz.ch, ETH Zürich, Switzerland; Kevin Koch, University of St. Gallen, Switzerland; Zimu Zhou, zimuzhou@smu.edu.sg, Singapore Management University, Singapore; Simon Föll, ETH Zürich, Switzerland; Xiaoxi He, ETH Zürich, Switzerland; Tina Menke, Karlsruhe Institute of Technology, Germany; Elgar Fleisch, ETH Zürich, Switzerland; Felix Wortmann, felix.wortmann@unisg.ch, University of St. Gallen, Switzerland.

2474-9567/2021/9-ART117 $15.00
https://doi.org/10.1145/3478078

**ACM Reference Format:**
Shu Liu, Kevin Koch, Zimu Zhou, Simon Föll, Xiaoxi He, Tina Menke, Elgar Fleisch, and Felix Wortmann. 2021. The Empathetic Car: Exploring Emotion Inference via Driver Behaviour and Traffic Context. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 3, Article 117 (September 2021), 34 pages. https://doi.org/10.1145/3478078

## 1 INTRODUCTION

The aim of affective computing is to recognise and adapt to the affective state of the user. Examples of its implications include reducing the user's frustration through adaptive and comfortable communication or delivering just-in-time interventions [66]. Driving is often a source of stress, and is associated with cognitive burden [74]. Traffic, driving task, and navigation systems, etc., frequently constitute triggers of negative emotions during driving [9, 36, 87] and hence lead to a sub-optimal mental status while driving than other daily tasks [9, 29]. The accumulated cognitive load and negative emotions do not only have negative consequences for the drivers' physiological well-being [74], but can also cause the immediate impairments of driving performance. For instance, *anger* and *sadness* are found to be associated with risky and degraded driving performance [8, 80] and a *positive valence* is found to be correlated with better steering behaviours [78]. It is of vital importance to detect negative or stress-related emotions, such as *anger*, *disgust*, *fear*, and *sadness*, as well as positive emotions such as *joy* and *valence*. In the environment of a car, technologies that can infer the driver's emotions can help improve their comfort, well-being, and safety. Thus, detecting the driver's emotional state is an important part of the vision for the car of the future. In pursuit of this vision, OEMs (*e.g.*, BMW, KIA and Mercedes–Benz [11, 32, 49]) and start-up companies (*e.g.*, Affectiva [48]) have already taken initial steps.

However, past research on in-vehicle emotion recognition has yielded two major limitations that hinder the large-scale integration of this technology into the car [4]. First, available solutions such as the ENERGIZING COACH, introduced by Mercedes–Benz [11], rely on smartwatches that incur additional cost, and hence are still not extensively used. Hence, such approaches are limited in their scalability. Second, existing work has focused on emotion recognition via facial expressions or speech, which particularly compromises drivers' privacy [88]. This substantially differs from the approach considered here because we rely on the control area network (CAN-bus) and front-view camera data. The CAN-bus is a standard for communication among in-vehicle sensors, controllers, and actuators, and contains detailed information about driving behaviours such as steering, braking, and accelerating. Front-view cameras can easily be mounted on vehicles to record videos that capture the ambient traffic environment, and are already built into the latest generation of cars. Hence, approaches based on CAN-bus and front-view camera can be applied to existing vehicles without requiring expensive, special-purpose hardware.

Our approach draws on an analogy between recent advancements in generic emotion recognition via *user behaviour* and *the contexts of application* to enable in-vehicle emotion inference. Many studies have shown the feasibility of inferring a users' emotions from the patterns of his/her smartphone usage [37, 42, 89]. Emotion-induced behavioural patterns are highly correlated with context (*e.g.*, dining, working, and entertainment), which can improve the accuracy of emotion recognition [23, 50]. Because driving is a unique activity with predefined rules and interactions that differ from what has been investigated in previous studies [37, 42, 89], we apply the concept of re-purposing the available sensor modalities to develop inference models targeting behavioural and context-based emotion recognition. In summary, the novelty of our approach is in applying machine learning to CAN-bus and front-view video data streams to reliably detect the emotions of drivers while minimising privacy-related concerns. Our approach is formalised within the following two research questions (RQ):

- *RQ1: To what extent can the emotions of drivers be inferred based on (a) CAN-bus data streams, (b) the front-view camera, and (c) a combination of both (fusion)?*
- *RQ2: How much improvement does emotion recognition based on vehicle data offer compared with state-of-the-art methods based on physiological sensors?*

In successfully answering these research questions, the main contributions and results of this study can be summarised as follows:

- We conducted a four-month field study involving nine participants to collect various (CAN-bus, front-view camera, driver facial camera and physiological sensors) empirical sensory data during uncontrolled daily driving activities on public roads. In total, we collected valid data on 675.6 hours driving data made by nine participants covering various scenarios.
- Recording CAN-bus and video data requires pre-processing in order to use them in machine learning pipelines. We outline a comprehensive pre-processing and feature engineering pipeline for both kinds of data. We comprehensively summarise important features for time-series and video data as the basis of the classification algorithms.
- We develop an emotion classification algorithm that can process and classify CAN-bus and video data streams as well as fuse them. Based on either kind of CAN-bus or video data, our algorithm can detect the emotions[1] based on facial expressions with an average macro F1-score of around 70% in user-dependent settings, and around 60% in user-independent settings. The results of our experiments showed that the fusion of the two modalities can further improve the performance.
- While the methods based on physiological sensors are the most prevalent among in-vehicle emotion recognition, there is a clear trend towards more ubiquitous affective state monitoring methods [88]. We demonstrate that our proposed method can accurately recognise drivers' emotions and achieve comparable performance as the medical-grade physiological sensor-based state-of-the-art baseline method [52]. Our solution is more ubiquitous, and uses only sensors available in modern cars.
- To the best of our knowledge, this is the first study that verifies the feasibility of non-intrusive inference of driver emotions in empirical situations based on driving behaviours and traffic contexts. Based on 675 hours of driving data collected on public roads in real driving scenarios, a challenging environment compared with laboratory conditions, our results are likely to be more reliable.

The remainder of this paper is organised as follows: We review related work in Section 2, and present our field study in Section 3. We introduce our methods for emotion recognition in Section 4. Section 5 summarises the results of verification of our method, and we discuss them further in Section 6. Finally, Section 7 provides the conclusions of our study.

## 2 RELATED WORK

In this section, we outline common standards of emotion measurement, the current trends in emotion recognition, and the progress in research on in-vehicle emotion recognition for drivers.

### 2.1 Emotions

To recognise the emotional state of drivers, it is important to obtain a reliable ground truth of emotions that can be used to train and evaluate models. The affective computing community often uses several expressions interchangeably to describe emotions [6], and there is no consensus on a general classification, even in the field of psychology [27]. The challenge is that the emotional spectrum ranges around different origins: short and raw (affect), directed and intensely felt (emotions), or long and diffuse (moods) [3, 69]. Researchers commonly summarise these differences in origins as experiences of feeling basic emotional states [69], and various models have been proposed to reliably measure these emotional states in a standardised way.

The common methods of measurement are discrete category models and two- (2D) or three-dimensional (3D) models [6]. Discrete category models (*e.g.*, [18]) allow subjects to categorise their emotional states into a set of basic emotions, such as happiness, sadness, anger, surprise, fear, and disgust. By contrast, 2D and 3D models

[1] In this work, we focused on: anger, disgust, fear, joy, neutral, sadness, surprise, and valence.

measure emotions in a multidimensional space [6]. An example is Russel's circumplex model, in which subjects can rate their levels of arousal (*i.e.*, degree of activeness) and valence (*i.e.*, degree of happiness) [62]. Combinations of the two express specific emotional states, *e.g.*, low arousal and low valence represent sadness, whereas high arousal and high valence indicate excitement.

## 2.2 Recent Advancements in Emotion Recognition

Inferring emotions is an objective that has been addressed in many prior studies. Although the task considered here is similar, the approach differs with regard to the input used. A variety of inputs, ranging from physiological sensors [52, 68] and facial images [40] to speech [20], have been used. Physiological sensors such as smartwatches allow for the continuous estimation of a subject's emotions, inferred based on the heart rate (variability), electrodermal activity, and accelerometer data. The potential of this technique has been recognised by researchers. Ubiquitous devices that record physiological data streams can be used to detect emotional states [52, 68]. However, monitoring physiological signals requires that subjects wear one or multiple devices, which may introduce inconvenience or discomfort for regular daily use. As an alternative, non-intrusive approaches have been developed. The most popular methods of emotion recognition are based on facial expressions and speech [20, 40, 81]. Nevertheless, the continuous recording of a user's visual or audio information may raise privacy-related concerns.

Inferring emotions from behaviours and contexts is a promising less-intrusive alternative. Data accumulated from a user's interactions with everyday devices contain behavioural and contextual information that can act as a proxy for the experiences or specific emotional states of the users. Researchers rely on devices such as smartphones to gather this information. The data gathered from smartphones are diverse, and contain information on app usage, screen time, accelerometer, GPS, SMS, call activity, Wi-Fi, and Bluetooth signals. These data constitute a digital representation of user behaviour and context, from which their emotional state can be deducted [79]. Several studies [5, 37, 59, 77, 89] have shown that users' emotions can be inferred from their patterns of mobile phone usage. Canzian *et al.* used mobility trace from a smartphone to detect a tendency toward depression [7]. The car, as an everyday device with sensor modalities, allows us to derive the driver's behaviour and context as the basis for our emotion recognition algorithms. We propose detecting the driver's emotions using driving behaviours as represented by the CAN-bus signals of the car and the context (surrounding traffic) as determined by the front-view video camera.

## 2.3 Facial Expressions and Their Annotations

Facial expressions are among the most informative source for the estimation of affective and cognitive states [17]. The Facial Action Coding System (FACS) is an objective and quantitative way to measure facial expression. In the FACS, *action units* describe the expressions currently active in the face at any given time, such as "brow furrow" and "eye widen". As a consequence, facial expressions can be quantified based on the combination and the level of presence of *action units* [19, 67]. Various existing works have relied on facial expressions or action units for the estimation of psychological states [31, 73, 91] or the detection of deception [71].

However, manual FACS-coding requires profound expert knowledge and the process is laborious due to the manual labelling required. With recent advances in computer vision and machine learning, numerous studies have proposed the automated recognition of facial expressions [2, 12, 46, 47, 84].

Developments in the automated recognition of facial expressions has had a major impact on affective computing. Whereas earlier works relied on the manual annotation of facial expressions, an increasing number of researchers now detect facial action units or acquire emotion labels by using various algorithms. For example, based on automatically detected facial action units, Sen *et al.* analysed deceptive communication [71], and Sharam *et al.* focused on the assessment of cognitive performance [73]. Rostaminia *et al.* leveraged the the output of detection of OpenFace [2] as ground truth labels for the unobtrusive sensing of upper facial action units [61]. To estimate emotional experiences during collaborative computer-aided design (CAD), Zhou *et al.* utilised the results of detection of facial expressions from Affectiva [47] as the emotion labels of CAD users [91].

Compared with self-report questionnaires, the automated annotation of emotions based on facial expressions has several advantages. First, automated annotation can significantly reduce the manual labour required, thus enabling the acquisition of a large number of emotion labels at a more temporally granular level. Second, the unobtrusive emotion annotation via facial expressions means that the subject's experience is uninterrupted. The frame-by-frame annotation of facial expressions enables dynamic representations of how emotion evolve over time [46]. Finally, by using facial expressions, the cognitive load imposed by self-reports is avoided and the subjects' responses are less likely to be biased due to the form of the questionnaires, their context, and other irrelevant factors [70].

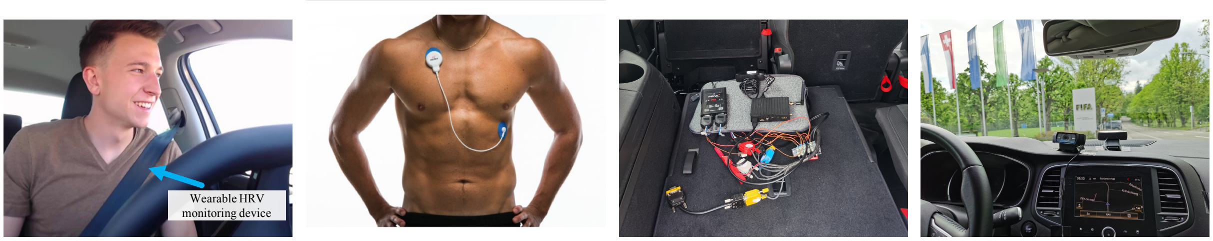
Given the above advantages, we acquire facial expressions-based emotion labels of drivers in this study by using a facial monitoring camera mounted on the dashboard of the vehicle. Past work [81] has shown that state-of-the-art algorithms can reliably detect the facial expressions of drivers with an accuracy of around 95% in various contexts. Thus, in this work, we rely on the automated facial expression annotation tool for the emotion label acquisition.

## 2.4 In-vehicle Emotion Recognition

As in the wider field of emotion recognition, researchers use sensors to detect the driver's emotions in cars.[2] Physiological sensors are preferred for measuring stress levels as a specific emotional response of the driver [26, 60, 64, 82] because stress and emotions in general are highly correlated with physiological measures, such as the heart rate, the variation in heart rate, and blood pressure [65]. Malta *et al.* used electrodermal activity (EDA) in combination with facial expressions, driving events, and pedal behaviours to build a Bayesian network to predict the frustration of drivers [43]. To infer the comprehensive mental and physical states (concentration, tension, tiredness, relaxation) of drivers, the authors of [57] built a body sensor network to monitor signals, the such as electrocardiogram (ECG), electroencephalography (EEG), electrodermal activity (EDA) and respiration rate. Kato *et al.* classified emotions as positive and negative based on ECG and pulse wave measurements during traffic jams [30]. Most in-vehicle emotion recognition based on physiological signals relies on numerous sensors, which are inconvenient to deploy. Data from physiological sensors as well as those on facial expressions were used by Zhang *et al.* to monitor a driver's emotional states and degree of fatigue [90]. Guang *et al.* introduced the first neuromorphic vision based distracted driving recognition system that analyses driver drowsiness, driver gaze-zone, driver hand-gesture behaviour from the generated streams of asynchronous events with a dynamic vision sensor [10]. Shafaei *et al.* proposed a multimodal system that combines facial expressions with steering wheel usage and vehicular acceleration for emotion recognition [72]. Facial expressions have been used in industry solutions in this vein. For example, Affectiva has developed an automotive software development kit that can analyse emotions using a driver-monitoring camera system [48]. Research has also used the driver's speech to detect emotions [76]. However, the fundamental barriers of emotion recognition also apply to these methods. The use of video data for the face or data for speech analysis raises privacy concerns. Physiological sensors are also particularly intrusive.

Given that today's car already have a set of sensor modalities, they are already well prepared for emotion inference. Cars are equipped with a large number of sensors that are accessible via the CAN-bus. In the CAN-bus, the sensors and actuators of a car transmit comprehensive information about driving-related activities and the vehicle's dynamics while the radar and camera systems interpret the environmental context. Researchers have shown that CAN-bus data can be used to detect driver behaviours to derive the relevant contextual information, for example, identifying the driver among a group of users [21], the profile of the driving style [44], and the

[2] We recommend a recent review by Zepf *et al.* [88] that reports details of research on emotion recognition in a vehicle.



(a) Participant in vehicle (b) Heart monitoring device [54] (c) CAN data collection device (d) Webcam deployment

Fig. 1. Experimental setup.

anticipation of the driver's intentions [25, 39]. Several studies have explored the use of CAN-bus information [13, 55] to detect stress using driving behaviours. As this past research indicates, the available sensor data allow for a wider interpretation beyond their intended usage (i.e., controlling the car). To the best of our knowledge, this is the first study to detect the driver's detailed emotional state by passively interpreting data streams available in today's cars.

## 3 DATA COLLECTION

Our analysis is based on a four-mouth field study during which a variety of empirical sensory data were collected from participants during daily drives. The field study involved nine participants (originally 10; data from one participant were corrupted, and were thus removed), and lasted from July 4 to November 5, 2019. Prior to the field study, we received approval from our university's ethics committee to conduct it.

### 3.1 Participants

We recruited nine (four females and five males, mean age, 37 ± 8 years) participants using an internal call in an enterprise with more than 1,000 employees. Our selection followed the idea of recruiting ordinary daily commuters. They were selected to represent a large variety of people (purposive sampling). Two participants were single and eight were married. Three had children and two had pets. The preferred activities while driving included making phone calls, listening to music or the radio, and talking to other occupants of the car. We assigned each participant the same type of vehicle (with modifications for data collection as described below). The participants were supposed to use the vehicles for their daily drives, including business trips and vacations.

### 3.2 Data Collection Equipment and Protocol

As mentioned in Section 1, our hypothesis is that the driver's emotion can be inferred from driving behaviours and traffic contexts, which can be measured in turn by the vehicle's CAN-bus and front-view cameras, respectively. The emotion labels based on ground-truth facial expressions were captured by another camera mounted on the dash-board of the vehicle. Because the state-of-the-art affect recognition schemes [26, 52, 68] rely on physiological sensors, we also collected physiological data of the participants for comparison.

We accessed the CAN-bus data via a PCAN-USB Pro FD-Adapter [24]. Two webcams (Logitech HD Pro Webcam C920) were mounted on the dashboard of the vehicle to record videos of the traffic context [3] as well as the driver's facial expressions. The CAN-bus and video data streams were controlled by an industrial-grade embedded

[3] We used a separate camera for this, and did not rely on the CAN-bus-based radar or camera systems, to make our analysis more flexible as both systems in the car had only a limited feature set available.

computer (Compulab IOT-GATE-IMX7), and were stored locally in the vehicles on external hard disks. When the vehicle was started, the computer initialised the recordings of both types of data. We collected 49 CAN signals, including those for the speed of the four wheels, accelerator position, angle of the steering wheel, and brake pedal pressure. A complete list of CAN signals is shown in Table 6 in Appendix A. Figure 1c and Figure 1d show setups for collecting the CAN-bus data, data from the front-view camera, and data from videos of the driver's face.

In line with previous emotion recognition studies that used physiological data [52, 68], we collected the heart rate (HR) and the heart rate variability (HRV) using a heart monitoring device (Firstbeat Bodyguard 2), as shown in Figures 1a and 1b. All participants were asked to wear the device for the two weeks of the field study. [4] The sampling rate of the HR and HRV of the heart monitoring device was 1000 $Hz$.

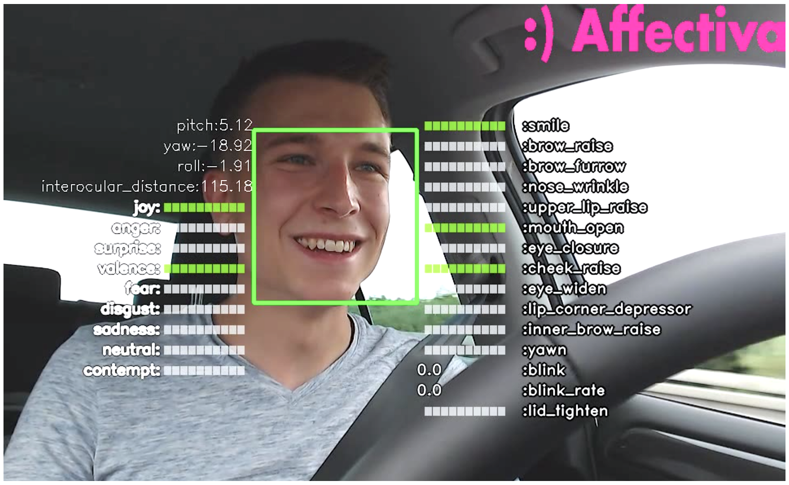


Fig. 2. An example of Affectiva annotation.

### 3.3 Characteristics of Driving Data

It was crucial for our dataset to capture representative driving situations. This subsection presents some important statistics related to our dataset.

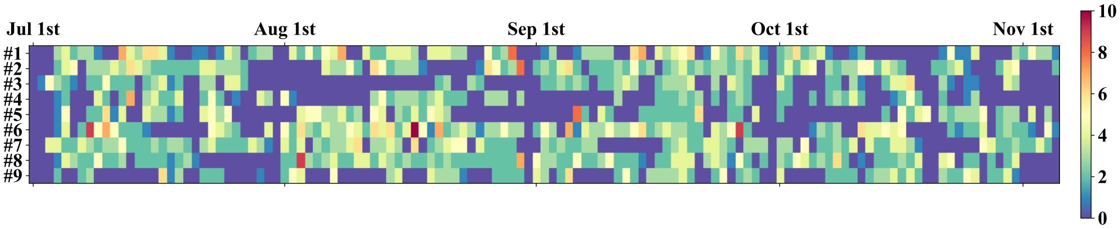
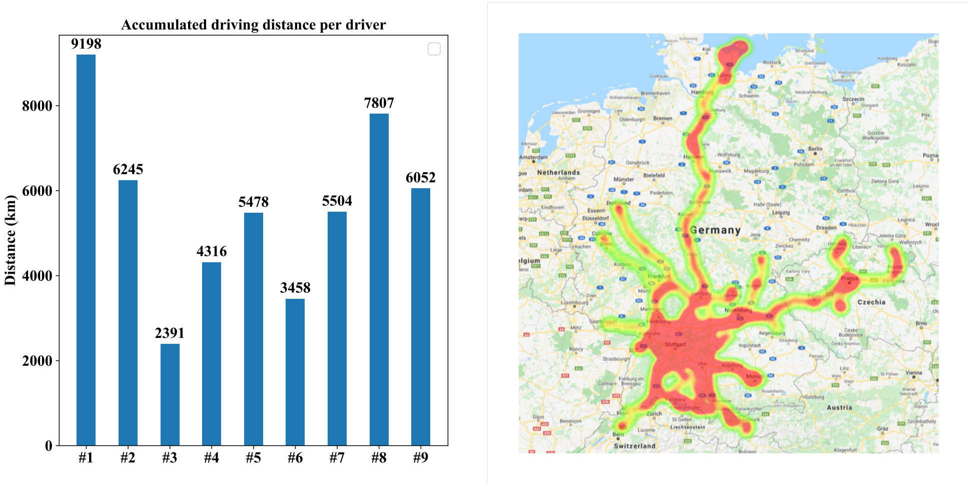


Fig. 3. Number of trips per driver and per day.

[4] We recorded the heart rate signals of the participants only in the first two weeks because wearing the heart monitoring device for a prolonged time may cause discomfort.



(a) Total driving distance per driver

(b) GPS heatmap of trips

Fig. 4. Spatial coverage of driving trips.

After data cleansing, we had around 675.6 hours of driving data with videos from front view camera as well as videos of the driver's face. Figure 3 plots the number of trips[5] per day for each participant during the field study. The overall average number of trips per day was greater than one, and the participants had up to 10 trips per day. Most participants were working during the period of our field study, except drivers #3 and #9, who had taken vacations in July and August. The total driving distances of each participant are plotted in Figure 4a. Most drivers drove for reasonably long distances (more than 3000 *km*) during the field study. The average trip length and duration were around 25 *km* and 23 *min*, respectively. The GPS records of the vehicles are presented as a heatmap in Figure 4b. As is shown there, most participants drove around the area of Stuttgart, Germany. A few long-distance trips were also taken to Prague and northern Germany.

Overall, our dataset covered a wide range of daily driving activities. All participants were active in terms of the number of trips per day and driving distance. The coverage and diversity of the dataset ensured that our experiments were generalisable to heterogeneous situations and drivers.

## 3.4 Characteristics of Emotion Labels

This subsection describes the statistics of the ground-truth emotion labels acquired from the Affectiva algorithm as well as the procedures for data cleansing, pre-processing, and transformation that were applied to them.

*3.4.1 Emotion Annotation Tool.* Affectiva spun out of MIT's Media Lab. Its emotion recognition technology uses computer vision algorithms and deep learning models to estimate emotions based on facial expressions. Unlike other solutions based on the recognition of facial expressions, Affectiva's algorithms are built on a very large foundational dataset, containing more than 9.7 million facial images of people from 90 countries, with over 5 billion facial frames and six years of video data. Affectiva's deep learning algorithms are optimised with automotive in-cabin data, including more than 20,000 hours featuring more than 4,000 unique individuals [48].

[5] A trip was defined as a segment of continual driving behaviour without a pause longer than 10 minutes.

Given these features, Affectiva's solution can reliably capture driver emotions. In our experiment, we used one of the latest stable versions (ics-2.2.1) of Affectiva for annotation.

Affectiva detects the facial expression of the subject for every frame. For all emotions except for *valence*, it outputs a score between 0 (absent) and 100 (present), indicating the *presence level* of the relevant emotions [47]. The emotions included anger, disgust, fear, joy, neutral, sadness, and surprise. The score of *valence* ranged from -100 to 100, and thus was divided into negative and positive valence (unhappy to happy).

*3.4.2 Emotion Persistence.* Facial expressions are often not long-lasting, with a duration between 0.5$s$ - 4$s$ [16]. An example is provided in Figure 5. To ensure that the driver's most prevalent and stable affective states were captured accurately, we applied a non-overlapping sliding window and divided the driving data into ***driving segments***. The emotion labels were defined according to their average level of presence in each ***driving segment***. Our objective was therefore to predict driver emotion using the CAN-Bus data and data from the front-view video of the same ***driving segments***. An illustration of this setting is provided in Figure 6. By adjusting the length of the sliding windows (and hence the length of the driving segments), driver emotion could be recognised at different granular levels. We defined the default length of the driving segments as 10 *mins* to capture emotions (directed and intensely) rather than affects (short and raw) or moods (long and diffuse) [6]. In addition, the combination of the sliding window and the temporally continuous annotation from Affectiva enabled emotion recognition at any time during drive.

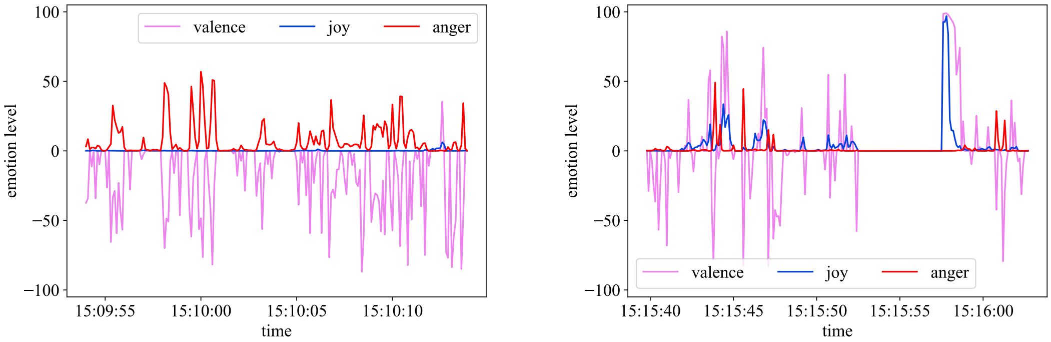
Owing to inevitable occlusion (e.g. from driver turning head), and undesirable illumination, facial expressions could not be detected in every frame. Only in a subset of the frames were both the driver's face and their facial expressions detected. We refer to such frames as *valid frames*. To ensure the quality of the label, we considered only the driving segments that contained more than 70% of valid frames. The labels of the driving segments were then computed as the average level of presence of an emotion over all valid frames. Such a quality check reduced the amount of driving data being used for training and testing, as some driving segments were discarded owing to an inadequate number of valid frames. From 675.6 hours of driving data, we obtained a total of $19,885$ two-minute driving segments (equivalent to 662.8 hours of data) or $3,377$ 10-minute driving segments (equivalent to 562.7 hours). Trips shorter than 10 *mins* were not included in the 10-minute driving segments, which led to the different number of driving hours between the types of segments.

*3.4.3 Emotion Distribution.* It is critical to inspect the label distribution to understand how reliable we can predict the emotions and to ensure that our ground truth is valid. As described in Section 3.4.2, the emotion labels were the average presence level of emotions over driving segments of a predefined length. Figure 7 illustrates the distribution of emotion labels of the 10-*min* driving segments The following observations can be made from it:

- The drivers elicited more negative *valence* than positive *valence*. This can be explained by the fact that driving is a task that requires high cognitive load and induces stress [9, 29].
- Driving had a varying impact on the drivers, which was reflected in the distinct personalised mean values of each emotion across different drivers.
- Most emotions have a low presence level. This can be explained by the instantaneous nature of facial expressions: The values shown in Figure 7 were averaged over segments of 10 *mins* of driving; owing to the instantaneous nature of facial expressions, as illustrated in Figure 5, the values were balanced out by non-present moments, in which a given expression was absent from the subject's face.

*3.4.4 Transformation of Emotion Labels .* Before conducting further data analysis, we needed to address common affective computing-related issues with our data.

[6] The analysis of different lengths of driving segments is provided in Section 5.3.2



(a) Dominance of anger and negative valence

(b) Dominance of joy and positive valence

Fig. 5. An example of the raw output of Affectiva results over time.

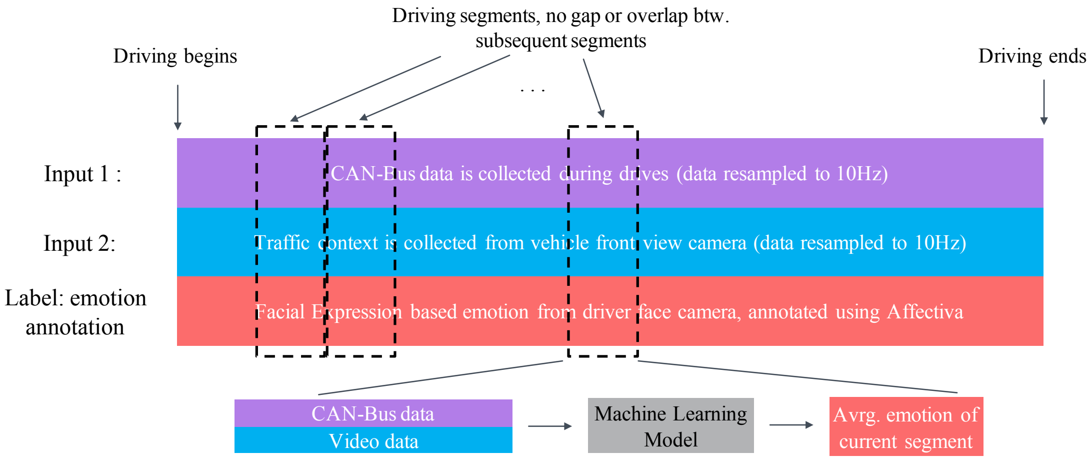

Driving segments, no gap or overlap btw.
subsequent segments
Driving begins
Driving ends
. . .
Input 1 :
CAN-Bus data is collected during drives (data resampled to 10Hz)
Input 2:
Traffic context is collected from vehicle front view camera (data resampled to 10Hz)
Label: emotion annotation
Facial Expression based emotion from driver face camera, annotated using Affectiva
CAN-Bus data
Video data
Machine Learning Model
Avrg. emotion of current segment

Fig. 6. Data and emotion labels used for our data analysis.

As discussed in Section 3.4.3, each participant had their personalised baseline (i.e. different mean values) of emotions in the context of driving because emotions are subjective, and their interpretation among people differs [45, 85, 86]. Such subjective factors introduced bias to the emotion recognition. Therefore, the emotion labels of each participant were calibrated following a personalised emotion label transformation, as described in Equation 1. This binarisation processing procedure was similar to that in [7, 14, 50, 71].

$$emotion\ label = \begin{cases} low\ class,\ if\ presence\ level < personalised\ median \\ high\ class,\ if\ presence\ level >= personalised\ median \end{cases} \tag{1}$$

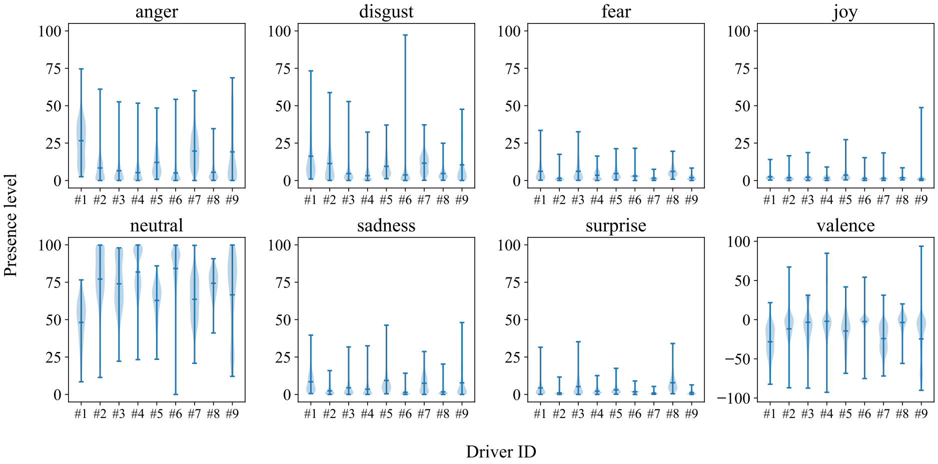


Fig. 7. Violin plots of the emotion distribution of each driver.

Such a transformation accounted for variations in individual perceptions of the driving task. Owing to the continuous values of the annotation, we have almost balanced the emotion labels for low and high classes (average proportion of majority class = 50.6% ± 0.7%) after label transformation. The objective of our final prediction was to determine whether a driver's emotion positively or negatively deviated from their personal baseline.

# 4 METHODOLOGY

This section explains our methodology to infer the driver's emotions based on their driving behaviours (CAN-bus data) and the traffic contexts (video data). We first introduce the data pre-processing and feature engineering for each sensor modality in Section 4.1, and then detail the inference models based on the CAN-bus and video data as well as their combination in Section 4.2.

## 4.1 Data Pre-processing and Feature Engineering

We briefly explain the candidate features extracted from CAN-bus, front-view video, and additional data sources. Note that we focus on interpretable features that have been proven to be effective in research on sensing the activities of the driver.

*4.1.1 Features from CAN-bus Data.* The CAN-bus data were used to capture driving behaviours and vehicular dynamics. By using 49 CAN data signals, we chose the following as candidate features because they were the most common across different vehicles and, thus, could guarantee the capability of generalisation of our analysis. They were as follows: angle of the steering wheel, yaw rate, brake pressure, pedal position of the accelerator, speeds of the four wheels, longitudinal and lateral acceleration, and rotational speed of the motor.

Because the recording of the raw CAN-bus data was not synchronised, we re-sampled them to 10 $Hz$, which is suitable for CAN-bus data processing [25]. Following the common practices for such data processing [21], the re-sampled CAN-bus data streams were split into sliding windows, from which features such as statistical features, auto-correlation etc., are derived to form a feature vector. Our tests of several sliding windows with different lengths resulted in five-second-long windows without overlap. These comparably short windows are common in CAN processing and seem to capture single driving manoeuvres [21, 39]. Table 1 lists the features derived from the CAN-bus data streams.

Table 1. Input signals and derived features from CAN-bus data. The features have been widely used for CAN-bus data processing [21], and were intended to capture driving behaviours and vehicular dynamics. The dimensions of certain features are noted in brackets.

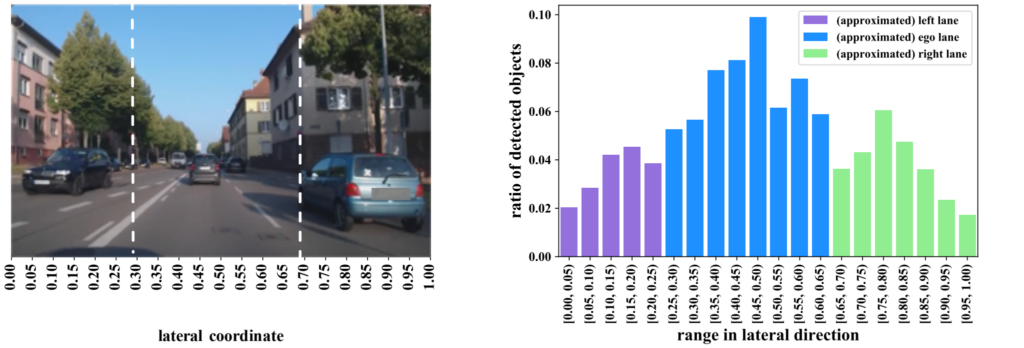
| Input Signal | Derived Features for CAN data |
|---|---|
| Steering wheel angle, | min, max, mean, |
| Yaw rate, | std. dev., median, |
| Brake pressure, | kurtosis, skewness, |
| Accelerator pedal angle, | quantile (25%, 75% and 95%), |
| Speeds of the four wheels, | piece-wise approximation (14D), |
| Longitudinal and lateral acceleration, | auto-correlation (50D), |
| Motor rotational speed | log(FFT) (26D), |

*4.1.2 Features of Front-view Video Data.* Video data from the front-view camera were expected to reflect the traffic contexts. Because we hypothesise that the driver's emotions can be inferred from traffic contexts, it is reasonable to assume that these contexts are easy to perceive and interpret by humans. Following this assumption, we derived the video-based features from objects detected by using the available object detection algorithms. Deriving pixel-wise features from videos for recognising the driver's emotion is beyond the scope of our work.

We applied Yolo-v3 [58] to detect a subset of objects related to the driving context (small vehicles, trucks, pedestrians, and cyclists), and used the location and size of each object as candidate features. The videos were re-sampled to 10 frames per second, and Yolo-v3 detection was used on each frame.

To infer driver emotion from traffic situation, it is important to involve lane information as it can affect the cognitive load on drivers [35]. With lane information, the relative position of the surrounding traffic participants can be better determined. However, state-of-the-art lane detection [53] algorithms are computationally expensive and limited in their availability to researchers. We used a simplified method to gain lane-related information. We split an image into three columns, corresponding to [0,0.25), [0.25,0.65) and [0.65,1] along the lateral axis. These columns were approximated as the left, middle, and right lane from a drivers perspective, respectively. In addition, past research [35] has shown that drivers are more sensitive to closer vehicles than distant ones. Therefore, the trade-off between accuracy and simplicity was acceptable even though the correspondence between the split and the lanes was valid only in the near-range of the ego-vehicle. Moreover, such an approximation was supported by the distribution of the detected positions and number of vehicles captured by the front-view camera as shown in Figure 8b. The lateral distribution of the detected objects was approximated by a Gaussian mixture model with three clustering centres. For each approximated lane we computed the statistical features and auto-correlation of the number and the sum of the sizes of the detected objects in it in each sliding window (the same sliding windows as for CAN-bus data). Each approximated lane was used to compute 120 features, for a total of 360 features from the results of detection using Yolo-v3. These are summarised in Table 2.

*4.1.3 Features from Auxiliary Data Sources.* Because emotions vary over the course of a day [15], we considered temporal features to recognize the emotions of the driver using the following: seconds before dawn, seconds after dusk, seconds before sunrise, seconds after sunset, indicator of driving at night, current time (formatted in the 24h-scale), and the day of the week. The first four features were set to zero if driving had occurred after or before



(a) Front view of webcam

(b) Lateral distribution of number of detected objects

Fig. 8. Simplified lane separation.

Table 2. Features of traffic.

| Input Signal | Derived Features for Video data |
|---|---|
| Yolo detection results:<br>- Class (vehicle, cyclist, or pedestrian)<br>- Confidence, coordinates of bounding boxes 10 most confidently detected objects in each frame | min, max, mean, std. dev., median, kurtosis, skewness, quantile (25%, 75%, or 95%), and auto-correlation of number of objects and sum of sizes in each approximated lane in a sliding window |

the corresponding event to ensure that there were no negative values in the temporal features. The temporal features were computed for every five seconds by using a sliding window based on the time associated with the corresponding windows. From each five-second sliding window, a temporal feature vector of seven dimensions was computed. That is to say, for instance, from a 10-*min* driving segment, a 120-*step* ($10min$ / $5s$ = 120) sequence of 7D temporal feature vectors was computed. We did not perform feature selection on the temporal features, and simply concatenated this temporal sequence of feature vectors to the sequences obtained from CAN data, videos, or a fusion of the two.

*4.1.4 Summary of Features.* From each sliding window, we computed 1100D, 360D, and 7D feature vectors for the CAN-bus, front-view video, and temporal modalities, respectively. Each modality therefore contained a sequence of feature vectors of the same sequence length. For example, the sequence length of a 10-*min* driving segment was 120 ($10min$ / $5s$ = 120). For each driver, we computed the p-values associated with each dimension of the feature vector by using an ordinary least-squared regression for every emotion. We selected only the 10 dimensions with the lowest p-values as input from each modality (i.e., the CAN-bus and front-view videos) per emotion for each driver. The cumulative distribution of the p-values of the selected dimensions of the feature vectors are plotted in Figure 15 in Appendix A. Note that for most signal sources, more than 80% of the selected features had p-values lower than 0.05. We leveraged multi-task learning approach, a method that has been proven to be useful in recognising emotions or activities [63, 68], to build a neural network to predict all emotions at the same time. This meant that for every driver in the training set, 80 (10D, eight emotions = 80) CAN features, 80 video features and seven temporal features were considered. If a feature was relevant to multiple emotions, it was selected only once. If multiple drivers were in the training set, the union of the selected features was used. Therefore, the number of selected features varied depending on the drivers in the training set.

## 4.2 Driving Behaviour- and Context-based Inference Models

In this subsection, we first introduce the driving behaviour- and context-based models using data from only either CAN-bus or front-view videos as input, followed by a combination of the two (*i.e.*, sensor fusion).

*4.2.1 CAN-bus-only Model.* Random forest approaches, as for example in [21], can be used to process CAN-bus data. However, they require hand-crafted features and are unable to capture temporal dependencies between these feature vectors. By contrast, the approaches based on recurrent neural networks (RNN) like the one in [25] can model the time dependence on a wider time scale even without explicit feature creation. However, such end-to-end training methods must learn the knowledge of carefully designed features and requires larger amounts of data. Our method combines the advantages of both. Our pipeline begins by computing the feature vectors of CAN-bus data using sliding windows as shown in Table 1 and Section 4.1.1. Then, the feature vectors are fed sequentially into a RNN.

The RNN can learn a high-level abstract summary of the data from the sequence of feature vectors. This summary is then processed by a fully connected network that outputs a probability distribution as a prediction of low and high states of a certain emotion. The detailed settings of the proposed method are as follows: We choose an RNN architecture with two layers, where each consists of 64 gated recurrent units (GRUs). This architecture is similar to [25], where they have input dimension of 665 and two layers RNN with 256 gated recurrent units is used. We have proportionally applied the similar reduction from our input dimension to the gated recurrent units. The CAN-bus and time feature vectors are concatenated in each sliding window. Our RNN has a simple structure because manual feature extraction and feature selection are applied a priori, which significantly reduces the complexity of the input. The input to the RNN is a sequence of feature vectors with reduced dimensions. We leveraged a multi-task learning approach to build a neural network that predicted all emotions at the same time. As the labels were almost balanced for every emotion, we randomly shuffled the training batches to obtain an almost equal number of low- and high-state samples on average for each emotion in every training batch. We used the Adam [33] optimizer and cross-entropy as loss functions, as described in Equation 2. The ReLU [51] was applied as activation function to the fully connected layers. The learning rate was set to 0.005. The hyper-parameters/parameters are empirically tuned to achieve the best emotion recognition performance. We trained the neural network until its loss converges. The inverse proportion to the class ratio was assigned as weight to the loss function:

$$\mathcal{L} = \sum_{i=1}^{n} -w_i[y_i \cdot log\ \sigma(x_i) + (1 - y_i) \cdot log(1 - \sigma(y_i))] \tag{2}$$

with

$$w_i = n/\sum_{j=1}^{n}(1 - y_i)(1 - y_j) + y_i y_j \tag{3}$$

where $x_i$ and $y_i$ are the prediction and ground truth for the $i^{th}$ sample, respectively, and n is the total number of samples. The framework of the proposed method is illustrated in the *CAN-only* branch of Figure 9.

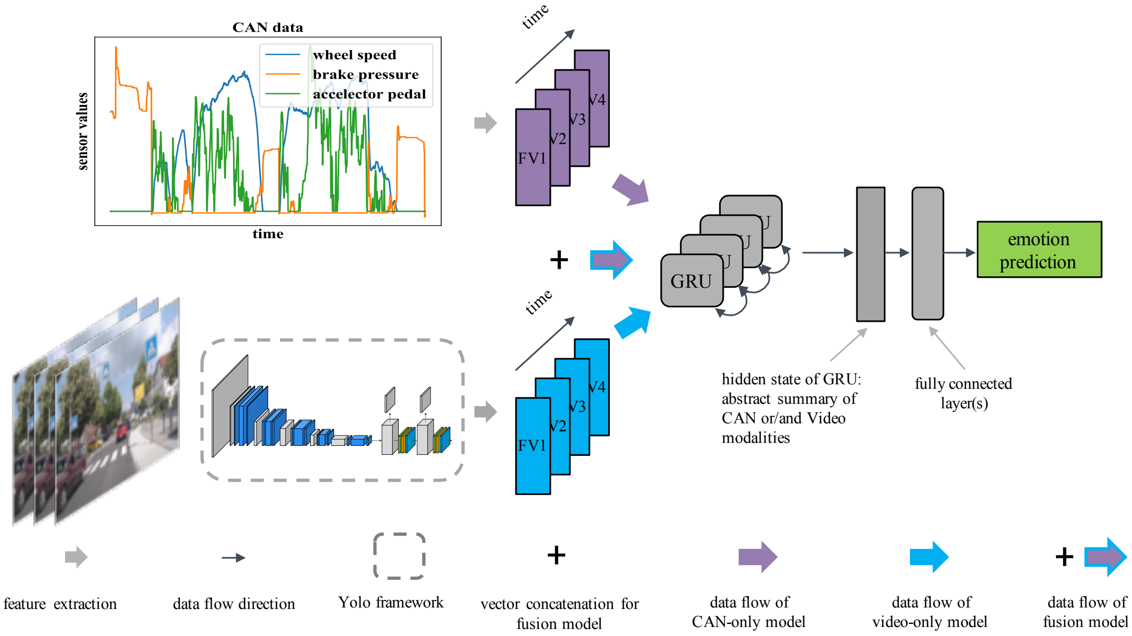


Fig. 9. Multi-modal recurrent architecture for recognising the driver's emotion: CAN-only model, Video-only model and fusion model.

*4.2.2 Front-view Video-only Model.* Apart from the modality of the input, the structure and settings for the front-view video-only model were identical to the CAN-bus-only model. The feature vectors of video replaced those of the CAN-bus as explained in Section 4.1.2. Vectors of the video and the time feature were concatenated in each sliding window. The framework of the front-view video-only model is illustrated in the *Video-only* branch of Figure 9.

*4.2.3 Fusion Model.* We then integrated the two sensor modalities into a joint inference model, called the fusion model. To fuse the sensor data of the CAN-bus with the front-view video, we constructed fused feature vectors formed by concatenating the feature vectors as described in Sections 4.2.1 and 4.2.2. The rest of the network architecture and settings constant are kept constant. The sensor fusion model is illustrated in the *fusion* branch of Figure 9.

## 5 EVALUATION

Our evaluation section has three parts. In Section 5.1, we evaluated our *driving behaviour- and context-based inference models*. In Section 5.2, we described the baseline methods and compared their results with that of our proposed method. Finally, in Section 5.3 we investigated the stability, usability, and complexity of our approach.

To avoid misinterpretation, the macro-F1-score that we used is formalised in Equations 4 - 7, where $tp$, $fp$, and $fn$ represent the true positive, false positive and false negative, respectively, depending on the level of presence of each emotion, and subscript $(c)$ distinguishes between low-and high-scoring classes, *i.e.*, $(c) = (low)\ or\ (high)$. For the sake of brevity, all *F1-scores* in this paper refer to the *macro F1-score*.

As mentioned in Section 3.4.2, the *driving segments* were generated by using non-overlapping sliding windows to ensure that there was no intersection between any pairs of *driving segments*. For intra-subject evaluation, we built a personalised model for each driver by randomly dividing their driving segments into training (70%) and test (30%) datasets. For the leave-one-subject-out (LOSO) evaluation, we used the driving segments from the $i^{th}$ driver as test data and the data from the remaining $N-1$ drivers is used as training data. In our case, $i$ was iterated from one to nine. All the experiments were repeated 10 times by using 10 different random seeds for both intra-subject and LOSO evaluations. The F1-scores presented are the average of the 10 repetitions of all drivers. The standard deviations of the 10 repetitions are indicated as error bars in the corresponding figures.

$$F1 = (F1_{(low)} + F1_{(high)})/2 \tag{4}$$

$$precision_{(c)} = tp_{(c)}/(tp_{(c)} + fp_{(c)}) \tag{5}$$

$$recall_{(c)} = tp_{(c)}/(tp_{(c)} + fn_{(c)}) \tag{6}$$

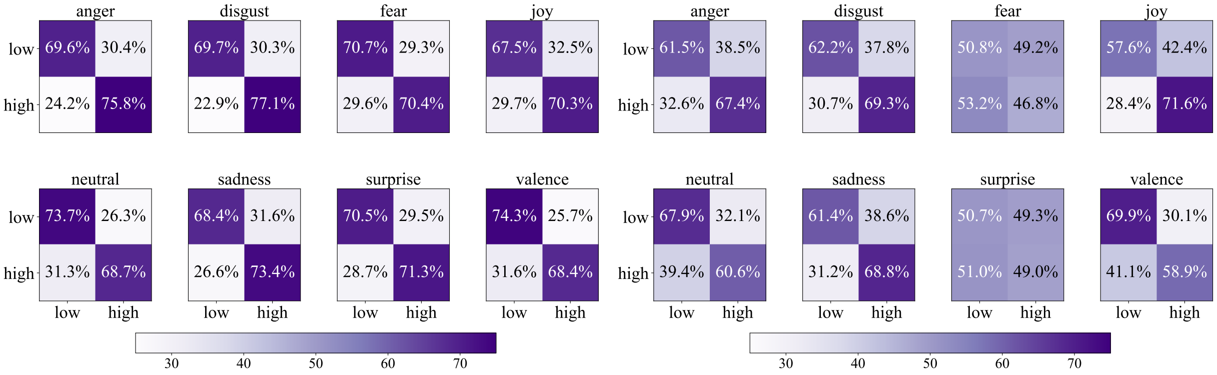
$$F1_{(c)} = \frac{2}{recall_{(c)}^{-1} + precision_{(c)}^{-1}} \tag{7}$$

### 5.1 RQ1: To what extent can the emotions of drivers be inferred based on (a) CAN-bus data streams, (b) front-view camera, and (c) the combination of both *(fusion)*?

Table 3 summarises the results of our proposed models for intra-subject and leave-one-subject-out (LOSO) evaluations. The best performance scores are highlighted in bold. In the intra-subject evaluation settings, the results indicate that the fusion model achieved the best performance scores on all emotions. For all emotions combined, our CAN-bus-only model achieved an F1-score of 68.8%. Our video-only and fusion models improved this score to 69.9% and 71.0%, respectively. With limited fluctuations, all emotions achieved comparable equally high scores. This suggests that a comprehensive recognition of the driver emotional state was possible in intra-subject setting. We then inspected the generalisability of the proposed models by exploring the results of the LOSO evaluation.

Compared with the setting for the intra-subject evaluation setting, the LOSO evaluation yielded lower scores. This was expected because emotion is subjective and user-independent emotion recognition remains a daunting challenge in affective computing community. The F1-scores achieved by the fusion model highlighted the varying performance across all emotions in the LOSO setting. Whereas the emotions of anger, disgust, neutral, sadness, and valence yielded relatively high scores of 63.8%, 64.5%, 62.8%, 63.5%, and 62.4%, respectively, those of fear, joy and surprise were less accurately detected, with F1-scores of 48.7%, 58.7% and 49.8%, respectively. These varying performance scores across emotions also accounted for the results of the CAN-bus-only and front-view video-only models. We conclude that some emotions were more difficult to detect than others independently of the sensor modality in the LOSO setting.

Furthermore, we plotted the confusion matrix of the fusion model to better visualise the results in Figure 10. For the personalised setting, the fusion model predicted the low and high classes with comparatively equal accuracy for all emotions, despite small variations. For the LOSO settings, we observed similar patterns (*i.e.*, higher values along the diagonal of the matrix) except in case of the emotions of *fear* and *surprise*. The confusion matrix shows that our proposed solution was not biased towards the low or the high class, which reflects the stability of our results. While the LOSO evaluation in general achieved lower F1-scores than the personalised evaluation, the confusion matrix showed that *anger*, *sadness*, and *valence* could be relatively accurately recognised in the LOSO settings. The more reliable detection of these three emotions in the LOSO settings is in line with existing findings



(a) Personalised model

(b) LOSO model

Fig. 10. Confusion matrix for fusion modality.

in literature providing evidence that they are closely related to driving performance and safety [8, 28, 78, 80]. For the sake of completeness, the confusion matrices of the CAN-only and video-only models are provided in Figures 16 and 17.

In summary, the F1-scores estimated using intra-subject and LOSO cross-validation indicate that emotions could best be inferred based on a combination of CAN-bus data and front-view camera data. The use of single modality models leads only to a minor reduction in performance.

Table 3. Intra-subject and LOSO cross-validation: comparison between the driving behavior- and context-based inference models as well as the fusion of both modalities. The best results of the three models (CAN, video, and fusion) are highlighted.

| | Personalised Model | | | LOSO Model | | |
|---|---|---|---|---|---|---|
| F1-score (%) | CAN | Video | Fusion | CAN | Video | Fusion |
| anger | 69.8 | 71.2 | ***72.3*** | 62.9 | 62.3 | ***63.4*** |
| disgust | 69.9 | 71.8 | ***72.8*** | 63.7 | 63.7 | **64.5** |
| fear | 69.1 | 69.2 | ***70.5*** | 48.0 | ***49.3*** | 48.7 |
| joy | 67.8 | 67.5 | ***68.6*** | ***59.3*** | 58.3 | 58.7 |
| neutral | 68.3 | 69.6 | ***70.9*** | 61.9 | 60.6 | ***62.8*** |
| sadness | 68.1 | 70.2 | ***70.6*** | 62.4 | 61.8 | ***63.5*** |
| surprise | 69.5 | 69.6 | ***70.9*** | 49.3 | 49.7 | ***49.8*** |
| valence | 68.3 | 70.4 | ***71.1*** | 60.5 | 60.2 | ***62.4*** |
| average | 68.8 | 69.9 | ***71.0*** | 58.5 | 58.2 | ***59.2*** |

### 5.2 RQ2: How much improvement does emotion recognition based on vehicle data offer compared with state-of-the-art methods based on physiological sensors?

The performance scores of our models were compared with a physiological signals-based baseline. Physiological signals are widely used for emotion recognition [13, 23, 26, 68] because emotions have an influence on the autonomic nervous system [34], which is highly correlated with such physiological measures as the heart rate, heart rate variability, and blood pressure [65]. We used the emotion recognition models proposed by Nardelli *et al.* [52] as our physiological sensor baseline. In [52], the emotion of subjects were induced by affective sounds in lab settings. During the experiments, the heart rate variability (HRV) data of the subjects were collected. The authors derived features of the time and frequency domains as well as non-linear features, and performed feature selection using the Friedman test and Wilcoxon signed-rank test. A quadratic discriminant classifier was then used for classification, and yielded an accuracy above 80% for *valence* and *arousal* by using the LOSO procedure. For the convenience of the readers, a comprehensive list of HRV features are provided in Table 7.

Combinations of *valence* and *arousal* express specific emotional states according to Russel's circumplex model [62]. For example, low arousal and low valence represent sadness, whereas high arousal and medium valence indicate happiness. Therefore, the model from [52], designed for valence and arousal recognition, was used as a proxy for the recognition of a variety of emotions, in preference to models that focus on detecting stress or frustration [26, 43, 60, 64]. Furthermore, unlike other physiological sensors-based approaches [23, 26, 68], the one in [52] does not require such physiological signals as the photo-plethysmogram, skin temperature, and skin conductance. Therefore, this baseline approach, by relying only on HRV data represents the available commercial wearable solutions to emotion recognition [41]. Moreover, our in-vehicle environment resembled the laboratory settings in [52] because the subjects were seated and no dramatic physical movements were allowed. Therefore, [52] was a suitable and competitive baseline.

We replicated the method developed by Nardelli *et al.* as our physiological sensor baseline, and refer to it as the *baseline* method. In accordance with our evaluation scheme in Figure 6, we used HRV data of the driving segments to predict the emotion labels in them. The baseline was evaluated and compared with our approach on a subset of the available driving data because HRV data were collected from the participants over only a two-week period. Therefore, only around 65 of the 675 hours of the data on the trips contained the relevant HRV data, corresponding to 388 samples of 10-*mins* driving segments. Although the HRV subset constituted only 9.6% of the entire dataset, it was still larger than the dataset used in [52], which consisted of only 208 samples.

Since the baseline [52] was conducted in lab condition, we optimised it to our context by exploring prevalent machine learning models in affective computing including Random Forest, Gradient Boosting, Support Vector Classifier, Extra Trees Classifier, Decision Trees as well as our proposed recurrent neural network. We performed grid search for optimal parameters and found that Extra Trees Classifier (the number of trees = 50, maximum depth of the tree = 20, and the minimal number of samples per split = 2) achieved the best performance. We refer to this optimised model as *baseline**.

A comparison between the proposed method and *baseline** method is provided in Table 4 [7]. Our proposed methods outperformed the *baseline** method in most of the cases. The scores in Table 4 are lower than those in Table 3. Such a decrease in model performance is expected, because we used a smaller dataset for comparison with the HRV baseline, and our proposed method relies on deep learning. We discuss this instance of performance and the dataset size in more detail in Section 5.3.1. The reduced performance of the baseline method compared to our approach is barely explainable due to a small dataset because our sample size (388) is much larger than the dataset (208) used in [52]. Therefore, our proposed method is better suited for in-vehicle emotion recognition than the state-of-the-art based physiological sensor-based approach.

## 5.3 Analysis of Stability, Usability, and Complexity of Proposed Solution

In this section, the stability, usability, and complexity of the proposed solution are analysed in detail.

*5.3.1 Stability vs. Dataset Size.* Like all intelligent learning systems, the performance of our proposed method is sensitive to the amount of training data used. Hence, we analysed changes in the performance of our models with

[7] The comparison between the proposed method and *baseline* method is provided in Table 8 in the Appendix

Table 4. Intra-subject and LOSO cross-validation: comparison between the baseline* and the proposed three models.

| | Personalised Model | | | | LOSO Model | | | |
|---|---|---|---|---|---|---|---|---|
| F1-score (%) | CAN | Video | Fusion | baseline* | CAN | Video | Fusion | baseline* |
| anger | 63.1 | 63.4 | 61.5 | 58.7 | 51.4 | 52.9 | 50.8 | 44.4 |
| disgust | 64.4 | 66.1 | 62.1 | 60.6 | 52.4 | 50.8 | 49.5 | 47.9 |
| fear | 62.0 | 59.5 | 56.8 | 55.8 | 54.7 | 54.0 | 50.6 | 49.9 |
| joy | 62.1 | 63.5 | 64.5 | 59.7 | 56.6 | 54.8 | 53.3 | 54.3 |
| neutral | 64.3 | 64.5 | 64.0 | 56.5 | 54.6 | 51.6 | 52.4 | 49.0 |
| sadness | 63.7 | 64.3 | 62.9 | 59.8 | 48.0 | 49.8 | 47.4 | 45.3 |
| surprise | 66.4 | 64.7 | 66.1 | 58.2 | 54.3 | 50.7 | 49.3 | 55.4 |
| valence | 66.7 | 61.0 | 62.4 | 62.5 | 47.9 | 50.2 | 46.4 | 45.7 |
| average | 64.1 | 63.4 | 62.5 | 59.0 | 52.5 | 51.8 | 50.0 | 49.0 |

the size of the training dataset for both intra-subject and LOSO evaluations. We changed the size of our training dataset by randomly removing a certain amount of data while keeping the size of the test dataset constant: a total of 30% of the dataset (intra-subject evaluation) or data for one driver (LOSO evaluation).

Figure 11a depicts the results of the stability analysis of the intra-subject evaluation. In general, despite fluctuations in case of small amounts of available data, the performance patterns of different modalities stabilised once the training dataset had reached 40% of its original size. After stabilisation, our fusion model outperformed the CAN-bus-only and front-view video-only models by around one standard deviation, independently of the size (> 40%) of the dataset. Overall, there is potential to further improve the performance by increasing training dataset size. However, it should be noted that the performance improvement relative to dataset size starts to show the sign of saturation after the training dataset size reaches 40% of its original size. Therefore, the performance we showed in Table 3 is very close the the upper bound if more data is available.

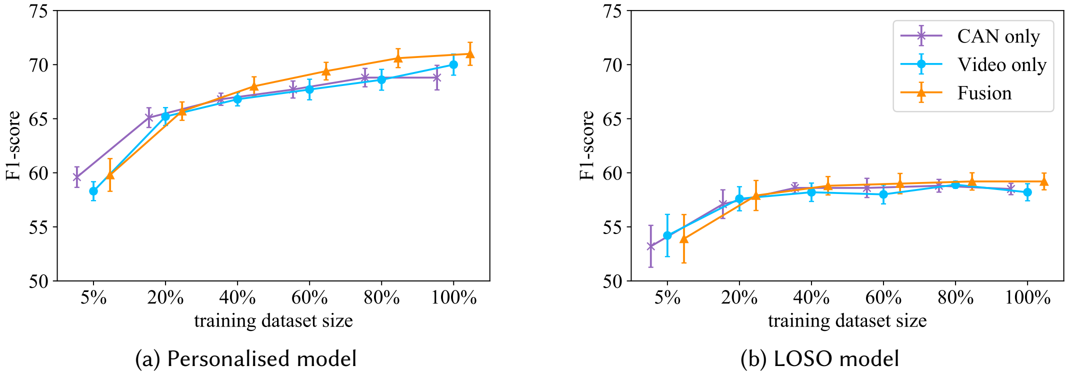
We performed the same stability analysis on the LOSO evaluation as shown in Figure 11b. The performance began to saturate as the size of the dataset increased to over 40% of the original dataset. After saturation, the fusion model consistently showed slight improvements over the CAN-only or Video-only models.

Another bottleneck in the LOSO model was the dataset diversity, *i.e.*, the number of training subjects. The higher the diversity of the dataset was, the greater was the possibility that the model finds generalisable patterns among the drivers. To verify this hypothesis, we evaluated our LOSO model with different numbers of subjects in the training dataset. This is depicted in Figure 12. While the overall performance saturated once four drivers had been included in the training dataset, the standard deviation dropped drastically as more drivers were included. This indicates the importance of diversity among the training subjects for model stability in emotion recognition.

*5.3.2 Analysis of Usability.* First, we analyse emotion recognition at different temporal granularities. As described in Section 3.4, the length of the driving segments could be adjusted to infer driver emotions at different granular levels. To inspect the performance of the model from this perspective, experiments were carried out on segments with lengths of 2-*mins* (19,884 samples), 5-*mins* (7,540 samples) and 10-*mins* (3,377 samples), and *entire segments* of trips [8] (1,773 samples).

Both the personalised and the LOSO models delivered the best performance on data on 10-*min* segments of driving. Despite having around two times more training samples than the setting for the 10-*min* driving segments, emotion recognition on the 5-*min* driving segments achieves inferior results. The performance on entire trips in

[8] A trip is defined as the duration from the driver starts driving until he/her ends driving. Therefore the trips have varying length.



(a) Personalised model

(b) LOSO model

Fig. 11. Model performance vs. size of training dataset.

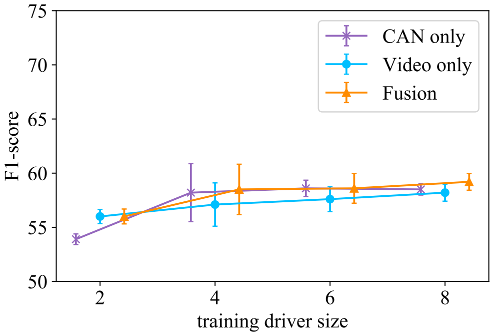
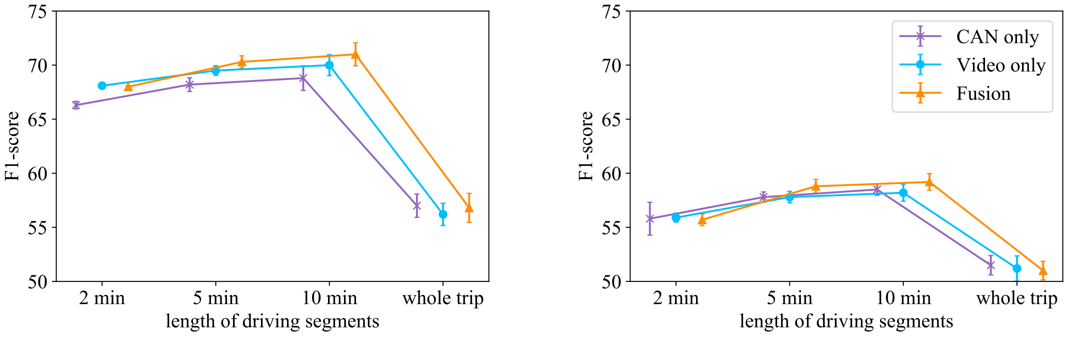


Fig. 12. Model performance vs. number of drivers in training set, LOSO model.

terms of recognition was by some distance the poorest. There are several possible explanations for this: a) driver emotions fluctuating while driving, where the proposed method was able to better infer a driver's prevalent emotional state over a certain period (in our case, around 10 *mins*) than the instantaneous detection of emotion, b) events during driving having varying impacts on the drivers' global emotions, and c) the total number of training samples for segments of entire trips being only half of those for the 10-*min* segments, in which case the reduced dataset had a negative impact on performance.

The proposed solution could best infer driver emotions at approximately 10-*min* intervals. Emotion recognition at a higher temporal resolution led to a decline in performance. This analysis shows that minute-level emotion recognition is possible by using the proposed solution. The high temporal resolution enables a more granular understanding of the evolution of driver's mental status evolving over time, and hence provides more opportunities for numerous applications. For example, just-in-time emotion regulation can be better applied at the appropriate time in this way.

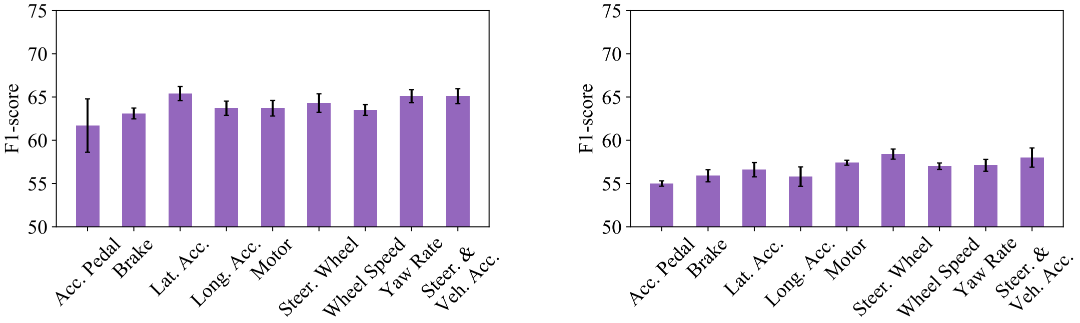
Next, the analysis of ablation study on CAN sensors is provided. We performed an ablation study by using only one, or a subset of, CAN sensor(s) to understand behaviour-based affective computing. The results are illustrated in Figure 14.



(a) Personalised model

(b) LOSO model

Fig. 13. Model performance vs. segment length.



(a) Personalised model

(b) LOSO model

Fig. 14. CAN-only model performance vs. ablation on CAN sensors: select only features from one sensor. Note: *Steer. & Veh. Acc* contains steer. wheel angle, lat. & long. acceleration and accelerator pedal; this subset replicates the settings in [72].

For personalised models, lateral acceleration (*lat. acc.*), *yaw rate*, and steering wheel angle and vehicle acceleration (*Steer. & Veh. Acc*) perform the best among all CAN sensors. That is to say, the emotion of individual person, to a certain degree, can be better traced by using vehicular dynamics (*i.e.*, *lat. acc.* and *yaw rate*) than the interaction between the driver and the vehicle (*accelerator pedal* and *brake*). For the LOSO model, steering wheel angle (*steer. wheel*) achieved the best F1-score with a low standard deviation, which means that steering wheel-related behaviours were more generalisable than other sensors among users.

Overall, the reduced number of CAN sensors had a negative impact on the *CAN-only model*. However, the degradation in performance in terms of emotion recognition performance was moderate. Even with only one CAN sensor (*lat. acc.* for personalised models and *steer. wheel* for LOSO models), the *CAN-only model* achieved F1-scores that were only around $2-3\%$ lower than if all sensors were used. The ablation study thus demonstrated the flexible usability of the proposed solution in case certain CAN sensors are not available.

*5.3.3 Model Complexity.* In this section, the model complexity of the proposed solution is analysed as it is vital for mobile and ubiquitous applications to run efficiently. Our model had a varying number of input dimensions depending on the feature selection process. Therefore, the number of model parameters was not fixed. As a reference, a model with a feature vector containing 100 dimension had 79,560 parameters and a size of 180KB. Time complexity was mainly evaluated based on the CPU setup: Intel Core i5 1.4 GHz Quad-Core, 16 GB LPDDR3. To better outline the performance of current advancements in GPU-accelerated parallel computing, Yolo-v3 object detection was also evaluated on GPU setup: Intel(R) Xeon(R) Silver 4110 CPU @ 2.10GHz, 196 GB LPDDR3, GeForce RTX 2080 Ti.

Table 5 shows the time complexity of each stage of processing of the proposed method on a 10 *min* segment. The proposed solution relied on explicit feature engineering in combination with a simple recurrent neural network. Therefore, the data pre-processing and the feature generation had higher time complexity. The main bottleneck of data pre-processing was the Yolo-v3 object detection on CPU-setup. However, potential improvements can help avoid this for ubiquitous in-vehicle emotion recognition, especially when the proposed solution is deployed on mobile devices:

- Subject to local legislative permission, the Yolo-v3 object detection can be delegated from mobiles devices to a cloud service with GPU devices. In our case, a 10-*min* segment of a video at 10FPS and a resolution of 640 × 360 had a size of only around 50 MB, which is acceptable in light of current mobile network capacity.
- An increasing number of mobile devices are now equipped with neural processing units that can significantly reduce the computation time and energy consumption for object detection on mobile devices.
- Our proposed video feature engineering relies only on size and location of the surrounding objects. Such information is already available in vehicles that are equipped with a radar or visual sensing systems. Thus, integrating the proposed solution into vehicles will not impose an additional computation burden.

Table 5. Time complexity for processing one 10 *min* driving segment.

| | CAN resampling | Yolo-v3 detection | | CAN feature generation | Video feature generation | Inference |
|---|---|---|---|---|---|---|
| | | CPU | GPU | | | |
| Wall-time (s) | 10.65 | 8028 | 75 | 1.22 | 1.05 | 0.20 |

In conclusion, the run-time complexity analysis demonstrates that the proposed solution has significant potential for the deployment on mobile devices (*e.g.*, in combination with driving recorders or smartphones mounted on dashboards), or for integration into an on-board computer of vehicles.

# 6 DISCUSSION

In this section, we discuss the results of our work. We outline the potential and limitations of the proposed non-intrusive approach with a special focus on privacy protection.

## 6.1 Non-intrusive In-Vehicle Emotion Recognition and its Contribution to the Car of the Future

In this subsection, we present the current solution offered by OEMs to infer emotions and highlight the possible impact of our approach on it. Braun *et al.* reviewed concept cars because corporate research is not often published in the technical literature [4]. For example, Audi presented its concept car, Elaine, which is similar to the solution proposed by Mercedes–Benz [1]. Elaine is capable of detecting stress and fatigue based on body temperature and heart rate obtained through wearable devices. Actions are subsequently triggered based on the recognised emotions. Such actions can be interventions, adaptive music, ambient light, or empathetic speech. Although the actions differ among Original Equipment Manufacturers (OEMs), most of them are similar in that emotions are recognised based on sensors that increase cost and might breach the user's privacy.

Unlike the above-mentioned emotion recognition methods that require physiological sensors or the facial images of users, our methods rely solely on a set of existing sensor modalities - CAN-bus and sensors that are going to be installed in future vehicles - traffic sensing systems, such as the radar, Lidar, and visual sensing systems. We used front-view cameras to capture traffic context. This paper showed that the analysis of driving behaviour and the context of an everyday task, i.e., driving, can help form a reliable estimate of the driver's emotional state. We used Affectiva to annotate the facial expressions of the drivers. Our algorithms were able to differentiate among 8 emotions to provide a detailed picture of the individually perceived emotions. In contrast to prevalent technologies that conduct frame-by-frame predictions (*e.g.*, Affectiva [48]), our solution offers an assessment of the emotion of the driver over a certain period, based on their driving behaviour and the traffic context. Although, initially it seems that both approaches are ambivalent, we believe that they are complementary. The emotional spectrum ranges around different origins. While Affetiva focuses on short and raw affect, our approach is more geared towards directed and less volatile emotions and can better protect user privacy.

Our approach is fully comparable with recent advancements in behaviour-based emotion recognition in the wild [5, 59, 77]. Since most in-vehicle emotion recognition relied on physiological sensors, facial expression, or speech and were conducted in lab settings [88], we benchmark our results with that of other in the wild studies based mainly on behaviours. Relying on Instagram photos, Reece and Danforth were able to detect the depression of users at an accuracy of 70% [59]. Taylor *et al.* built a personalised model to classify binary states (sad/happy) at an accuracy of 78.4%; similarly, Buda *et al.* [5] predicted lower or upper outliers of users' happiness and achieved and macro F1-score at 64.7% and 59.2% [9] for user-dependent and user-independent model, respectively. In particular, [77] excluded neutral days from both train and test, and [5] focused on classification of anomaly. In comparison, we did not exclude any data in our approach, meaning that our method shows more robustness in predicting ambiguous emotion responses that are close to the boundary between low and high classes. As such, our competitive performance against the state-of-the-art approaches demonstrate that our in-vehicle solution can serve as a good complement of prevalent behaviour-based emotion recognition approaches.

Our proposed non-intrusive system to recognise driver emotions allows for and encourages new opportunities to exploit the available data streams to infer the emotional state of the driver. Because our data were collected on public roads in empirical driving scenarios, a challenging environment was provided compared with laboratory conditions, and so our results are likely to be more reliable.

## 6.2 Emotion Recognition with Respect to Privacy Protection

We evaluated our method in terms of protecting user privacy. CAN-bus data and the surrounding visual information are much less sensitive to intrusion or leaking than physiological and facial data. For example, physiological data can reveal sensitive health conditions. Capturing surrounding traffics via a camera can violate privacy laws in certain regions (*e.g.*, the General Data Protection Regulation (GDPR) in Europe [22]). In such regions, the proposed *CAN-only model* can be applied as it achieves similar performance to that of the *Video-only* and *fusion models* while better ensuring preserve user privacy. Nevertheless, our models require only locations and distances (we used object size as a proxy) of the surrounding traffic participants. Surrounding object information can be retrieved from on-board radar systems [75]. Acquiring such information does not capture sensitive information (*e.g.*, license plates, facial images or explicit activities) from traffic participants. Hence, our approach adheres to the idea of privacy by design by minimising the data collection. It is compliant with current privacy-related laws and should be more acceptable to customers than prevalent technologies, such as driver-monitoring cameras

[9] Both macro F1-scores are computed from their reported confusion matrix and data distribution.

or health-sensor-based approaches. Thus, the novelty of our non-invasive emotion recognition algorithms is in their potential for scalability and privacy-preserving capabilities.

### 6.3 Flexible Deployment

The proposed *CAN-only*, *Video-only* and *Fusion models* delivered similar performances, as described in Section 5.1. We thus believe that the proposed methods have great flexibility in terms of deployment depending on the available sensor modalities:

- The *CAN-only model* can be deployed in regions in which vehicular front-view videos are disallowed (by law), or in the vehicles that are not equipped with the requisite sensing systems to capture traffic context information. Furthermore, the *CAN-only model* is robust against adversarial attacks and can be used in security sensitive scenarios.
- The *Video-only model* is more flexible than *CAN-only model*. On the one hand, in the regions (*e.g.*, China, India and USA) where the front-view video are permitted, the *Video-only model* can be used in combination with driving recorders or smartphones mounted on the vehicle dashboard. Such a setting allows for the recognition of driver emotion without access to CAN data, and hence further increases the ubiquity. On the other hand, the *Video-only model* is a natural fit in the context of increasingly intelligent vehicles. Vehicles in the future will have better sensing capability of surrounding traffic. The proposed *video-only model* relies exactly on the location and the distance of surrounding traffic participants. Such information can be reused in more and more intelligent vehicles.
- The *fusion model* further improved emotion recognition performance compared with single modality models. It can be applied to achieve the best user experience if the relevant conditions allow for it.

Our comparison should inform future research in the area on the suitability of these three models. Hence, we believe that our findings actually revealed high flexibility of the proposed method in the deployment.

### 6.4 Limitations

Despite our best efforts, this study has several limitations. First, our data collection took place in real-world traffic, a complex environment. To ensure a competitive baseline, we relied on heart rate variability measures obtained by FirstBeat - a recording device for cardiovascular activity. To record the data, electrodes needed to be attached to the chests of the subjects. Such a procedure is cumbersome, and requires additional training for the subject to correctly attach the electrodes. Moreover, our prototypical setup, its cost, and the cumbersome physiological sensors forced us to limit our sample size to nine drivers. These limitations should be, however, evaluated under the consideration that ours is the first study to explore drivers' emotions inference in a longitudinal setup (over four months). Our comprehensive sensor set that was used to collect information on driver behaviour and the environment of the car covered a wide range of influential factors.

Furthermore, we did not explicitly analyse the relation between emotion and specific driving manoeuvres or traffic events. Such a relation was modelled implicitly by using a neural network. Identifying driving manoeuvres or traffic events remains challenging. Even the state-of-the-art methods can identify only simple manoeuvres, such as *lane changing* or *turning right while accelerating* [38, 56, 83]. Future research should address this issue and a more explainable model is promised to better benefit both academia and industries.

Lastly, the LOSO model does not achieve as good performance as personalised model. Generalising from a user-dependent to a user-independent model remains a challenging topic in affective computing community. However, our results show that a generilisability across users is possible, especially for the emotions such as *anger*, *sadness*, and *valence* that are very relevant to driving safety. Further research is needed to investigate the extent to which such generalisation is achievable.

## 7 CONCLUSION

In-vehicle emotion recognition can enable applications of intelligent automobiles to improve comfort, well-being, and safety by adapting the car to the needs of the drivers. Current applications rely on physiological, facial, and speech-based data for emotion recognition. However, physiological sensors are cumbersome to wear during daily commute, and incur additional costs. Moreover, recording facial expressions and speech may raise privacy-related concerns. In this paper, we leverage recent advancements in generic emotion recognition through user behaviours, and used the idea for in-vehicle emotion inference. We relied solely on data streams of today's cars (*i.e.*, CAN-bus and front-view camera data): a strong advantage that allows for a scalable and privacy-preserving implementation in cars. We collected four months of CAN-bus front-view video data from nine users under naturalistic driving settings on public roads. The in-situ emotions of the drivers' faces were recorded by a monitoring camera and the relevant videos were annotated by using a the state-of-the-art facial expression recognition software - Affectiva.

Our results can be summarised as follows: First, we evaluated our models based on intra-subject and LOSO evaluations, and compared the performance of different sensor modalities. This evaluation revealed that a fusion model that combined CAN-bus data and front-view video data achieved the best results among our models. A single modality model yielded similar performance scores to those of a fusion model, which allows for the flexible deployment of the proposed model when certain sensor modalities are unavailable. Second, we compared the results of the proposed model with an HRV baseline. This evaluation revealed that our models can achieve comparable performance as the HRV baseline. Therefore, inferring driver emotions within a vehicle based on driving behaviour and context by using data from CAN-bus and video segments yields better ubiquity than physiological sensor-based approaches.

## ACKNOWLEDGEMENTS

This research was funded and supported by the Bosch IoT Lab at the ETH Zürich and University of St. Gallen.

## REFERENCES

[1] Audi AG. 2018. Audi Elaine. https://www.audi.com/en/experience-audi/models-and-technology/concept-cars/audi-elaine.html Accessed: 2021-04-23.

[2] T. Baltrusaitis, P. Robinson, and L. Morency. 2016. OpenFace: An open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 1–10.

[3] Lisa Feldman Barrett. 2006. Are Emotions Natural Kinds? *Perspectives on Psychological Science* 1, 1 (2006), 28–58.

[4] Michael Braun, Florian Weber, and Florian Alt. 2020. Affective Automotive User Interfaces – Reviewing the State of Emotion Regulation in the Car. arXiv:cs.HC/2003.13731

[5] Teodora Sandra Buda, Mohammed Khwaja, and Aleksandar Matic. 2021. Outliers in Smartphone Sensor Data Reveal Outliers in Daily Happiness. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1, Article 5 (March 2021), 19 pages.

[6] Rafael A. Calvo and Sidney D'Mello. 2010. Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications. *IEEE Transactions on Affective Computing* 1, 1 (2010), 18–37.

[7] Luca Canzian and Mirco Musolesi. 2015. Trajectories of Depression: Unobtrusive Monitoring of Depressive States by Means of Smartphone Mobility Traces Analysis. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, New York, NY, USA, 1293–1304.

[8] Michelle Chan and Anthony Singhal. 2015. Emotion matters: Implications for distracted driving. *Safety Science* 72 (2015), 302–309.

[9] Kiron Chatterjee, Samuel Chng, Ben Clark, Adrian Davis, Jonas De Vos, Dick Ettema, Susan Handy, Adam Martin, and Louise Reardon. 2020. Commuting and wellbeing: a critical overview of the literature with implications for policy and future research. *Transport Reviews* 40, 1 (2020), 5–34.

[10] Guang Chen, Fa Wang, Weijun Li, Lin Hong, Jörg Conradt, Jieneng Chen, Zhenyan Zhang, Yiwen Lu, and Alois Knoll. 2020. NeuroIV: Neuromorphic Vision Meets Intelligent Vehicle Towards Safe Driving With a New Database and Baseline Evaluations. *IEEE Transactions on Intelligent Transportation Systems* (2020), 1–13.

[11] Stephen Corby. 2017. Mercedes-Benz Vitality Coach revealed. https://www.drive.com.au/motor-news/mercedes-wants-to-make-you-fitter--even-while-you-re-sitting-still-gvkm25 Accessed: 2020-05-10.

[12] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. 2011. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*. 2106–2112.

[13] Chelsea Dobbins and Stephen Fairclough. 2019. Signal Processing of Multimodal Mobile Lifelogging Data towards Detecting Stress in Real-world Driving. *IEEE Transactions on Mobile Computing* 18, 3 (2019), 632–644.

[14] B. Egilmez, E. Poyraz, Wenting Zhou, G. Memik, P. Dinda, and N. Alshurafa. 2017. UStress: Understanding college student subjective stress using wrist-based passive sensing. In *IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. IEEE, Piscataway, NJ, USA, 673–678.

[15] Boris Egloff, Anja Tausch, Carl-Walter Kohlmann, and Heinz Walter Krohne. 1995. Relationships between time of day, day of the week, and positive mood: Exploring the role of the mood measure. *Motivation and Emotion* 19, 2 (1995), 99–110.

[16] P. Ekman. 2007. *Emotions Revealed: Recognizing Faces and Feelings to Improve Communication and Emotional Life.* Weidenfeld & Nicolson Ltd, London, U.K.

[17] Paul Ekman, Wallace V Friesen, and Sonia Ancoli. 1980. Facial signs of emotional experience. *Journal of Personality and Social Psychology* 39 (1980), 1125–1134.

[18] Paul Ekman, Wallace V Friesen, Maureen O'sullivan, Anthony Chan, Irene Diacoyanni-Tarlatzis, Karl Heider, Rainer Krause, William Ayhan LeCompte, Tom Pitcairn, Pio E Ricci-Bitti, Klaus Scherer, and Athanase Tzavaras. 1987. Universals and Cultural Differences in the Judgments of Facial Expressions of Emotion. *Journal of Personality and Social Psychology* 53 (1987), 712–717.

[19] P. Ekman and E. Rosenberg. 1997. *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System.* Oxford University Press, New York.

[20] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray. 2011. Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases. *Pattern Recognition* 44, 3 (2011), 572–587.

[21] Miro Enev, Alex Takakuwa, Karl Koscher, and Tadayoshi Kohno. 2016. Automobile Driver Fingerprinting. *Proceedings on Privacy Enhancing Technologies* 1 (2016), 34–50.

[22] European Union. [n.d.]. General Data Protection Regulation (GDPR) Compliance Guidelines. https://gdpr.eu/ Accessed: 2020-09-24.

[23] Martin Gjoreski, Hristijan Gjoreski, Mitja Luštrek, and Matjaž Gams. 2016. Continuous Stress Detection using a Wrist Device: in Laboratory and Real Life. In *Proceedings of the International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct.* ACM, New York, NY, USA, 1185–1193.

[24] PEAK-System Technik GmbH. 2019. PCAN-USB Pro FD-Adapter. https://www.peak-system.com/PCAN-USB-Pro-FD.366.0.html Accessed: 2020-05-10.

[25] David Hallac, Suvrat Bhooshan, Michael Chen, Kacem Abida, and Jure Leskovec. 2018. Drive2vec: Multiscale state-space embedding of vehicular sensor data. In *Proceedings of the IEEE International Conference on Intelligent Transportation Systems*. IEEE, Piscataway, NJ, USA, 3233–3238.

[26] Jennifer A. Healey and Rosalind W. Picard. 2005. Detecting Stress During Real-world Driving Tasks using Physiological Sensors. *IEEE Transactions on Intelligent Transportation Systems* 6, 2 (2005), 156–166.

[27] Carroll E. Izard. 2009. Emotion Theory and Research: Highlights, Unanswered Questions, and Emerging Issues. *Annual Review of Psychology* 60, 1 (2009), 1–25.

[28] Myounghoon Jeon. 2016. Don't Cry While You're Driving: Sad Driving Is as Bad as Angry Driving. *International Journal of Human–Computer Interaction* 32, 10 (2016), 777–790.

[29] D. Kahneman, A. B. Krueger, D. A. Schkade, Schwarzm N., and A. A. Stone. 2004. A survey method for characterizing daily life experience: the day reconstruction method. *Science* 306(5702) (2004), 1776–1780.

[30] Tomokazu Kato, Haruki Kawanaka, Md. Shoaib Bhuiyan, and Koji Oguri. 2011. Classification of positive and negative emotion evoked by traffic jam based on electrocardiogram (ECG) and Pulse wave. In *2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC)*. 1217–1222.

[31] C. D. Katsis, N. Katertsidis, G. Ganiatsas, and D. I. Fotiadis. 2008. Toward Emotion Recognition in Car-Racing Drivers: A Biosignal Processing Approach. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 38, 3 (2008), 502–512.

[32] KIA. 2019. Amplify Your Joy with Emotive Driving. https://pr.kia.com/en/future/future/emotive-driving-ces.do Accessed: 2020-09-01.

[33] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. http://arxiv.org/abs/1412.6980

[34] Sylvia D. Kreibig. 2010. Autonomic nervous system activity in emotion: A review. *Biological Psychology* 84, 3 (2010), 394–421.

[35] Yi-Ching Lee, John D. Lee, and Linda Ng Boyle. 2007. Visual Attention in Driving: The Effects of Cognitive Load and Visual Disruption. *Human Factors* 49, 4 (2007), 721–733.

[36] Alexander Legrain, Naveen Eluru, and Ahmed M. El-Geneidy. 2015. Am stressed, must travel: The relationship between mode choice and commuting stress. *Transportation Research Part F: Traffic Psychology and Behaviour* 34 (2015), 141–151.

[37] Robert LiKamWa, Yunxin Liu, Nicholas D. Lane, and Lin Zhong. 2013. MoodScope: Building a Mood Sensor from Smartphone Usage Patterns. In *Proceedings of the Annual International Conference on Mobile Systems, Applications, and Services.* ACM, New York, NY, USA, 389–402.

[38] HaiLong Liu, Tadahiro Taniguchi, Yusuke Tanaka, Kazuhito Takenaka, and Takashi Bando. 2017. Visualization of Driving Behavior Based on Hidden Feature Extraction by Using Deep Learning. *IEEE Transactions on Intelligent Transportation Systems* 18, 9 (2017), 2477–2489.

[39] Shu Liu, Kevin Koch, Bernhard Gahr, and Felix Wortmann. 2019. Brake Maneuver Prediction – An Inference Leveraging RNN Focus on Sensor Confidence. In *Proceedings of the IEEE International Conference on Intelligent Transportation Systems*. IEEE, Piscataway, NJ, USA, 3249–3255.

[40] André Teixeira Lopes, Edilson de Aguiar, Alberto F De Souza, and Thiago Oliveira-Santos. 2017. Facial Expression Recognition with Convolutional Neural Networks: Coping with Few Data and the Training Sample Order. *Pattern Recognition* 61 (2017), 610–628.

[41] Garmin Ltd. 2019. What Is the Stress Level Feature on My Garmin Watch? https://support.garmin.com/en-US/?faq=WT9BmhjacO4ZpxbCc0EKn9 Accessed: 2021-04-23.

[42] Li Lu, Jiadi Yu, Yingying Chen, Yanmin Zhu, Minglu Li, and Xiangyu Xu. 2019. I3: Sensing Scrolling Human-computer Interactions for Intelligent Interest Inference on Smartphones. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 97:1–97:22.

[43] Lucas Malta, Chiyomi Miyajima, Norihide Kitaoka, and Kazuya Takeda. 2011. Analysis of Real-World Driver's Frustration. *IEEE Transactions on Intelligent Transportation Systems* 12, 1 (2011), 109–118.

[44] Clara Marina Martinez, Mira Heucke, Fei-Yue Wang, Bo Gao, and Dongpu Cao. 2017. Driving Style Recognition for Intelligent Vehicle Control and Advanced Driver Assistance: A Survey. *IEEE Transactions on Intelligent Transportation Systems* 19, 3 (2017), 666–676.

[45] H. P. Martínez, G. N. Yannakakis, and J. Hallam. 2014. Don't Classify Ratings of Affect; Rank Them! *IEEE Transactions on Affective Computing* 5, 3 (2014), 314–326.

[46] D. McDuff, R. E. Kaliouby, and R. W. Picard. 2012. Crowdsourcing Facial Responses to Online Videos. *IEEE Transactions on Affective Computing* 3, 4 (2012), 456–468.

[47] Daniel McDuff, Abdelrahman Mahmoud, Mohammad Mavadati, May Amr, Jay Turcot, and Rana el Kaliouby. 2016. AFFDEX SDK: A Cross-Platform Real-Time Multi-Face Expression Recognition Toolkit. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, New York, NY, USA, 3723–3726.

[48] A. Mcmanus. 2020. Affectiva Automotive AI. https://go.affectiva.com/auto Accessed: 2020-10-28.

[49] A. Mcmanus. 2020. BMW: How In-Cabin Sensing Helps Build the Ultimate In-Vehicle Experience. https://blog.affectiva.com/bmw-how-in-cabin-sensing-helps-build-the-ultimate-in-vehicle-experience Accessed: 2020-10-28.

[50] Varun Mishra, Tian Hao, Si Sun, Kimberly N. Walter, Marion J. Ball, Ching-Hua Chen, and Xinxin Zhu. 2018. Investigating the Role of Context in Perceived Stress Detection in the Wild. In *Proceedings of ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*. ACM, New York, NY, USA, 1708––1716.

[51] Vinod Nair and Geoffrey E. Hinton. 2010. Rectified Linear Units Improve Restricted Boltzmann Machines. In *International Conference on International Conference on Machine Learning*. Omnipress, Madison, WI, USA, 807–814.

[52] Mimma Nardelli, Gaetano Valenza, Alberto Greco, Antonio Lanata, and Enzo Pasquale Scilingo. 2015. Recognizing Emotions Induced by Affective Sounds through Heart Rate Variability. *IEEE Transactions on Affective Computing* 6, 4 (2015), 385–394.

[53] D. Neven, B. D. Brabandere, S. Georgoulis, M. Proesmans, and L. V. Gool. 2018. Towards End-to-End Lane Detection: an Instance Segmentation Approach. In *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, Piscataway, NJ, USA, 286–291.

[54] Firstbeat Technologies Oy. 2019. Firstbeat Bodyguard 2. https://international-shop.firstbeat.com/product/bodyguard-2/ Accessed: 2020-05-10.

[55] Pablo E. Paredes, Francisco Ordonez, Wendy Ju, and James A. Landay. 2018. Fast & Furious: Detecting Stress with a Car Steering Wheel. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 665–677.

[56] Xishuai Peng, Yi Lu Murphey, Ruirui Liu, and Yuanxiang Li. 2020. Driving maneuver early detection via sequence learning from vehicle signals and video images. *Pattern Recognition* 103 (2020), 107276.

[57] Genaro Rebolledo-Mendez, Angélica Reyes, Sebastian Paszkowicz, Mari Carmen Domingo, and Lee Skrypchuk. 2014. Developing a Body Sensor Network to Detect Emotions During Driving. *IEEE Transactions on Intelligent Transportation Systems* 15, 4 (2014), 1850–1854.

[58] Joseph Redmon and Ali Farhadi. 2018. Yolov3: An incremental improvement. arXiv:1804.02767

[59] Andrew G. Reece and Christopher M. Danforth. 2017. Instagram photos reveal predictive markers of depression. *EPJ Data Science* 6 (2017).

[60] G. Rigas, Y. Goletsis, and D. I. Fotiadis. 2012. Real-Time Driver's Stress Event Detection. *IEEE Transactions on Intelligent Transportation Systems* 13, 1 (2012), 221–234.

[61] Soha Rostaminia, Alexander Lamson, Subhransu Maji, Tauhidur Rahman, and Deepak Ganesan. 2019. W!NCE: Unobtrusive Sensing of Upper Facial Action Units with EOG-Based Eyewear. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 1, Article 23 (March 2019), 26 pages.

[62] James A. Russell. 1980. A Circumplex Model of Affect. *Journal of Personality and Social Psychology* 39, 6 (1980), 1161–1178.

[63] Aaqib Saeed, Tanir Ozcelebi, and Johan Lukkien. 2019. Multi-Task Self-Supervised Learning for Human Activity Detection. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 2, Article 61 (June 2019), 30 pages.

[64] Aaqib Saeed and Stojan Trajanovski. 2017. Personalized driver stress detection with multi-task neural networks using physiological signals. In *Machine Learning for Health Workshop at 31st Conference on Neural Information Processing Systems*. Curran Associates, Inc.

[65] Akane Sano and Rosalind W. Picard. 2013. Stress Recognition using Wearable Sensors and Mobile Phones. In *Proceedings of the Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE, Piscataway, NJ, USA, 671–676.

[66] Hillol Sarker, Moushumi Sharmin, Amin Ahsan Ali, Md. Mahbubur Rahman, Rummana Bari, Syed Monowar Hossain, and Santosh Kumar. 2014. Assessing the Availability of Users to Engage in Just-in-Time Intervention in the Natural Environment. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Seattle, Washington) *(UbiComp '14)*. Association for Computing Machinery, New York, NY, USA, 909–920.

[67] M. A. Sayette, J. F. Cohn, J. M. Wertz, M. A. Perrott, and D. J. Parrott. 2001. A Psychometric Evaluation of the Facial Action Coding System for Assessing Spontaneous Expression. *Journal of Nonverbal Behaviour* 25, 3 (2001), 167–185.

[68] Philip Schmidt, Robert Dürichen, Attila Reiss, Kristof Van Laerhoven, and Thomas Plötz. 2019. Multi-target Affect Detection in the Wild: an Exploratory Study. In *Proceedings of the ACM International Symposium on Wearable Computers*. ACM, New York, NY, USA, 211–219.

[69] Norbert Schwarz. 1990. *Feelings as information: Informational and motivational functions of affective states.* Guilford, New York, NY, USA, 527–561.

[70] N. Schwarz and F. Strack. 1999. Reports of Subjective Well-Being: Judgmental Processes and Their Methodological Implications. *Well-Being: The Foundations of Hedonic Psychology* (1999), 61–84.

[71] Taylan Sen, Md Kamrul Hasan, Zach Teicher, and Mohammed Ehsan Hoque. 2018. Automated Dyadic Data Recorder (ADDR) Framework and Analysis of Facial Cues in Deceptive Communication. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4, Article 163 (Jan. 2018), 22 pages.

[72] Sina Shafaei, Tahir Hacizade, and Alois Knoll. 2019. Integration of Driver Behavior into Emotion Recognition Systems: A Preliminary Study on Steering Wheel and Vehicle Acceleration. In *Computer Vision – ACCV 2018 Workshops*. Springer International Publishing, Cham, 386–401.

[73] Kshitij Sharma, Evangelos Niforatos, Michail Giannakos, and Vassilis Kostakos. 2020. Assessing Cognitive Performance Using Physiological and Facial Features: Generalizing across Contexts. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3, Article 95 (Sept. 2020), 41 pages.

[74] Alois Stutzer and Bruno S. Frey. 2008. Stress that Doesn't Pay: The Commuting Paradox. *The Scandinavian Journal of Economics* 110, 2 (2008), 339–366.

[75] Yuliang Sun, Tai Fei, and Nils Pohl. 2019. A High-Resolution Framework for Range-Doppler Frequency Estimation in Automotive Radar Systems. *IEEE Sensors Journal* 19, 23 (2019), 11346–11358.

[76] A. Tawari and M. M. Trivedi. 2010. Speech Emotion Analysis in Noisy Real-World Environment. In *20th International Conference on Pattern Recognition*. IEEE, Piscataway, NJ, USA, 4605–4608.

[77] Sara Taylor, Natasha Jaques, Ehimwenma Nosakhare, Akane Sano, and Rosalind Picard. 2020. Personalized Multitask Learning for Predicting Tomorrow's Mood, Stress, and Health. *IEEE Transactions on Affective Computing* 11, 2 (2020), 200–213.

[78] Lana M. Trick, Seneca Brandigampola, and James T. Enns. 2012. How fleeting emotions affect hazard perception and steering while driving: The impact of image arousal and valence. *Accident Analysis & Prevention* 45 (2012), 222–229.

[79] Alina Trifan, Maryse Oliveira, and José Luís Oliveira. 2019. Passive Sensing of Health Outcomes Through Smartphones: Systematic Review of Current Solutions and Possible Limitations. *JMIR Mhealth Uhealth* 7, 8 (2019).

[80] Geoffrey Underwood, Peter Chapman, Sharon Wright, and David Crundall. 1999. Anger while driving. *Transportation Research Part F: Traffic Psychology and Behaviour* 2, 1 (1999), 55–68.

[81] Bindu Verma and Ayesha Choudhary. 2018. A Framework for Driver Emotion Recognition using Deep Learning and Grassmann Manifolds. In *Proceedings of the IEEE International Conference on Intelligent Transportation Systems*. IEEE, Piscataway, NJ, USA, 1421–1426.

[82] Jeen-Shing Wang, Che-Wei Lin, and Ya-Ting C. Yang. 2013. A k-nearest-neighbor classifier with heart rate variability feature-based transformation algorithm for driving stress recognition. *Neurocomputing* 116 (2013), 136–143.

[83] Jie Xie, Allaa R. Hilal, and Dana Kulić. 2018. Driving Maneuver Classification: A Comparison of Feature Extraction Methods. *IEEE Sensors Journal* 18, 12 (2018), 4777–4784.

[84] Huiyuan Yang, Umur Ciftci, and Lijun Yin. 2018. Facial Expression Recognition by De-Expression Residue Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[85] Y. Yang and H. H. Chen. 2011. Ranking-Based Emotion Recognition for Music Organization and Retrieval. *IEEE Transactions on Audio, Speech, and Language Processing* 19, 4 (2011), 762–774.

[86] G. N. Yannakakis, R. Cowie, and C. Busso. 2017. The ordinal nature of emotions. In *International Conference on Affective Computing and Intelligent Interaction (ACII)*. 248–255.

[87] Sebastian Zepf, Monique Dittrich, Javier Hernandez, and Alexander Schmitt. 2019. Towards empathetic car interfaces: emotional triggers while driving. In *Extended Abstracts of SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, Article LBW0129, 6 pages.

[88] Sebastian Zepf, Javier Hernandez, Alexander Schmitt, Wolfgang Minker, and Rosalind W. Picard. 2020. Driver Emotion Recognition for Intelligent Vehicles: A Survey. *ACM Computer Survey* 53, 3 (2020), Article 64.

[89] Xiao Zhang, Wenzhong Li, Xu Chen, and Sanglu Lu. 2018. MoodExplorer: towards Compound Emotion Detection via Smartphone Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (2018), 176:1–176:30.

[90] Y. Zhang, M. Chen, N. Guizani, D. Wu, and V. C. M. Leung. 2017. SOVCAN: Safety-Oriented Vehicular Controller Area Network. *IEEE Communications Magazine* 55, 8 (2017), 94–99.

[91] Jinxuan (Janice) Zhou, Vrushank Phadnis, and Alison Olechowski. 2020. Analysis of Designer Emotions in Collaborative and Traditional Computer-Aided Design. *Journal of Mechanical Design* 143, 2 (08 2020).

# A APPENDIX

Table 6. All signals of CAN data.

| Signal ID | Description | Signal ID | Description |
|---|---|---|---|
| 1 | Accelerator pedal position | 26 | Left turn indicator |
| 2 | Belt buckle indicator 1 | 27 | Longitudinal acceleration |
| 3 | Belt buckle indicator 2 | 28 | Motor rotational speed |
| 4 | Belt buckle indicator 3 | 29 | Odometer |
| 5 | Belt buckle indicator 4 | 30 | Parking light indicator |
| 6 | Belt buckle indicator 5 | 31 | Rear fog light |
| 7 | Brake indicator | 32 | Right turn indicator |
| 8 | Brake pressure | 33 | Steering wheel angle |
| 9 | Clutch switch | 34 | Steering wheel direction |
| 10 | Daytime running lamp | 35 | Steering wheel velocity |
| 11 | Dimmed headlights indicator | 36 | Steering wheel velocity direction |
| 12 | Electronic stability control | 37 | Tank level percent |
| 13 | External temperature sensor 1 | 38 | Temperature sensor |
| 14 | External temperature sensor 2 | 39 | Time |
| 15 | Flasher | 40 | Wheel direction (back left) |
| 16 | Fog light indicator | 41 | Wheel direction (back right) |
| 17 | Front wiper | 42 | Wheel direction (front left) |
| 18 | Gear position | 43 | Wheel direction (front right) |
| 19 | GPS altitude coordinate | 44 | Wheel speed (back left) |
| 20 | GPS latitude coordinate | 45 | Wheel speed (back right) |
| 21 | GPS longitude coordinate | 46 | Wheel speed (front left) |
| 22 | High beam | 47 | Wheel speed (front right) |
| 23 | High beam indicator | 48 | Yaw rate |
| 24 | Humidity | 49 | Yaw rate direction |
| 25 | Lateral acceleration | | |

Table 7. HRV features of the baseline method. [52]

| Feature Index | Description |
|---|---|
| | ***Time Domain Measures*** |
| 1 | the mean value (RR mean: R refers to the peak of the electrocardiography wave; RR is the interval between successive Rs) |
| 2 | the standard deviation (RR std) |
| 3 | the standard deviation of Normal-to-Normal (NN) intervals (SDNN) |
| 4 | the square root of the mean of the sum of the squares of differences between subsequent NN intervals (RMSSD) |
| 5 | the number of successive differences of intervals which differ by more than 50 ms, expressed as a percentage of the total number of heartbeats analyzed (pNN50) |
| 6 | the integral of the probability density distribution divided by the maximum of the probability density distribution (HRV tiangular index) |
| 7 | the triangular interpolation of NN interval histogram (TINN) |
| | ***Frequency Domain Measures*** |
| 8-10 | the power calculated within the very low frequency (VLF), low frequency (LF), and high frequency (HF) bands |
| 11-13 | the frequencies containing maximum magnitude (VLF peak, LF peak, and HF peak). |
| 14-16 | the power expressed as percentage of the total power (VLF power %, LF power %, and HF power %) |
| 17-18 | the power normalized to the sum of the LF and HF power (LF power nu and HF power nu) |
| 19 | the LF/HF power ratio |
| | ***Nonlinear HRV Measures*** |
| 20 | Approximate Entropy |
| 21-22 | Detrended Fluctuation Analysis: short-term fluctuations ($\alpha_1$) and long-term flunctuations ($\alpha_2$) |
| | Lagged Poincaré Plots: SD1, SD2, SD12, S, SDRR (details below) |
| 23 | SD1: the standard deviation related to the points that are perpendicular to the line-of-identity |
| 24 | SD2: the standard deviation that describes the long-term dynamics and measures the dispersion of the points along the identity line. |
| 25 | SD12 (SD1/SD2): the ratio between SD1 and SD2. |
| 26 | S ($\pi$SD1SD2): the area of an imaginary ellipse with axes SD1 and SD2 |
| 27 | SDRR: an approximate relation indicating the variance of the whole HRV series |

Table 8. Intra-subject and LOSO cross-validation: comparison between the baseline and our driving behaviour- and context-based inference.

| | Personalised Model | | | | LOSO Model | | | |
|---|---|---|---|---|---|---|---|---|
| F1-score (%) | CAN | Video | Fusion | baseline | CAN | Video | Fusion | baseline |
| anger | 63.1 | 63.4 | 61.5 | 54.0 | 51.4 | 52.9 | 50.8 | 43.6 |
| disgust | 64.4 | 66.1 | 62.1 | 55.7 | 52.4 | 50.8 | 49.5 | 44.3 |
| fear | 62.0 | 59.5 | 56.8 | 55.2 | 54.7 | 54.0 | 50.6 | 53.3 |
| joy | 62.1 | 63.5 | 64.5 | 61.4 | 56.6 | 54.8 | 53.3 | 54.5 |
| neutral | 64.3 | 64.5 | 64.0 | 58.3 | 54.6 | 51.6 | 52.4 | 44.9 |
| sadness | 63.7 | 64.3 | 62.9 | 54.1 | 48.0 | 49.8 | 47.4 | 44.4 |
| surprise | 66.4 | 64.7 | 66.1 | 53.9 | 54.3 | 50.7 | 49.3 | 49.2 |
| valence | 66.7 | 61.0 | 62.4 | 56.5 | 47.9 | 50.2 | 46.4 | 48.7 |
| average | 64.1 | 63.4 | 62.5 | 56.1 | 52.5 | 51.8 | 50.0 | 47.9 |

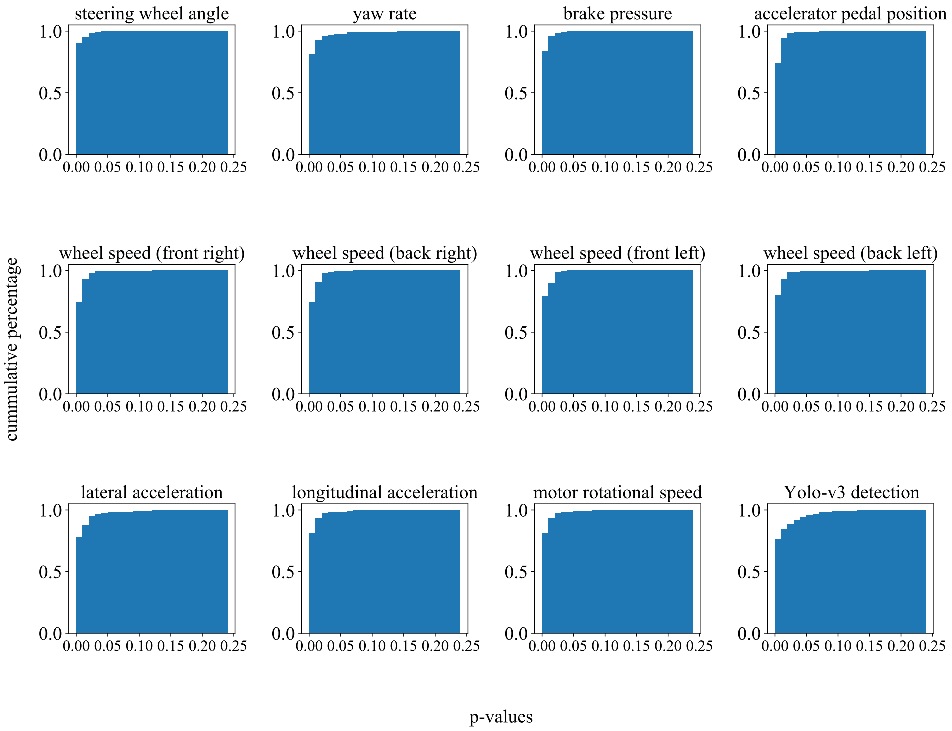
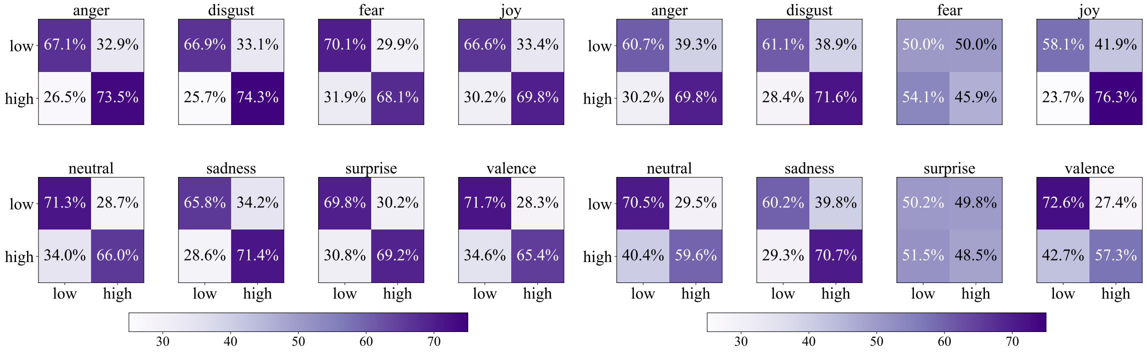


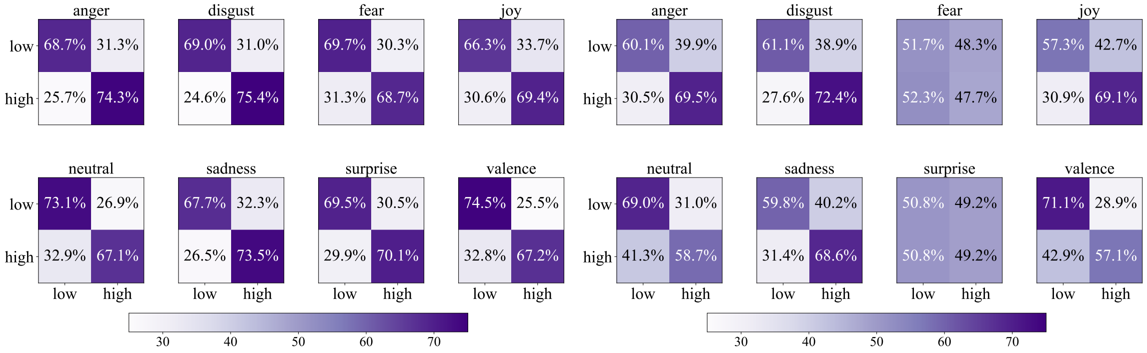
Fig. 15. Cumulative distribution of p-values of the selected features according to the source of the signal.



(a) Personalised model

(b) LOSO model

Fig. 16. Confusion matrix for CAN-only modality.



(a) Personalised model

(b) LOSO model

Fig. 17. Confusion matrix for Video-only modality.

Table 9. Results of precision for low class: Intra-subject and LOSO cross-validation: comparison between the driving behaviour- and context-based inference models as well as the fusion of both modalities.

| low class | Personalised Model | | | LOSO Model | | |
|---|---|---|---|---|---|---|
| precision (%) | CAN | Video | Fusion | CAN | Video | Fusion |
| anger | 71.6 | 73.5 | 76.6 | 64.1 | 64.3 | 66.7 |
| disgust | 71.4 | 73.7 | 74.1 | 64.7 | 65.3 | 65.0 |
| fear | 67.9 | 68.7 | 69.0 | 45.9 | 47.2 | 45.6 |
| joy | 68.7 | 67.4 | 71.0 | 60.2 | 58.0 | 61.1 |
| neutral | 64.0 | 65.5 | 68.5 | 53.9 | 53.2 | 58.6 |
| sadness | 71.4 | 73.7 | 74.0 | 65.6 | 64.8 | 66.8 |
| surprise | 70.2 | 68.6 | 70.2 | 49.4 | 47.6 | 47.9 |
| valence | 65.4 | 65.2 | 69.1 | 52.1 | 51.0 | 57.2 |
| average | 68.8 | 69.5 | 71.6 | 57.0 | 56.4 | 58.6 |

Table 10. Results of precision for high class: Intra-subject and LOSO cross-validation: comparison between the driving behaviour- and context-based inference models as well as the fusion of both modalities.

| high class | Personalised Model | | | LOSO Model | | |
|---|---|---|---|---|---|---|
| precision (%) | CAN | Video | Fusion | CAN | Video | Fusion |
| anger | 67.5 | 68.7 | 68.1 | 58.3 | 56.8 | 58.8 |
| disgust | 67.5 | 69.5 | 70.7 | 58.8 | 57.8 | 62.0 |
| fear | 70.3 | 69.7 | 71.9 | 50.5 | 51.1 | 52.3 |
| joy | 66.2 | 67 | 65.8 | 48.5 | 51.5 | 48.8 |
| neutral | 72.6 | 73.4 | 73.0 | 66.3 | 63.7 | 64.9 |
| sadness | 64.3 | 66.4 | 67.0 | 55.1 | 55.5 | 58.0 |
| surprise | 68.8 | 70.2 | 71.6 | 49.1 | 51.4 | 51.8 |
| valence | 71.3 | 76.1 | 73.0 | 63.9 | 65.8 | 65.3 |
| average | 68.6 | 70.1 | 70.1 | 56.3 | 56.7 | 57.7 |

Table 11. Results of recall for low class: Intra-subject and LOSO cross-validation: comparison between the driving behaviour- and context-based inference models as well as the fusion of both modalities.

| low class | Personalised Model | | | LOSO Model | | |
|---|---|---|---|---|---|---|
| recall (%) | CAN | Video | Fusion | CAN | Video | Fusion |
| anger | 66.4 | 68.3 | 69.2 | 60.0 | 59.7 | 61.1 |
| disgust | 66.6 | 68.9 | 69.8 | 60.8 | 61.0 | 62.3 |
| fear | 69.5 | 70.3 | 71.2 | 49.4 | 52.3 | 51.3 |
| joy | 67.3 | 65.8 | 67.7 | 58.8 | 56.8 | 57.8 |
| neutral | 71.6 | 72.9 | 73.5 | 70.8 | 68.8 | 67.7 |
| sadness | 66.4 | 67.9 | 69.2 | 60.8 | 60.0 | 62.2 |
| surprise | 69.8 | 70.2 | 71.2 | 50.2 | 51.5 | 51.4 |
| valence | 72.3 | 75.3 | 74.7 | 73.2 | 71.9 | 70.3 |
| average | 68.7 | 70.0 | 70.8 | 60.5 | 60.3 | 60.5 |

Table 12. Results of recall for high class: Intra-subject and LOSO cross-validation: comparison between the driving behaviour- and context-based inference models as well as the fusion of both modalities.

| high class | Personalised Model | | | LOSO Model | | |
|---|---|---|---|---|---|---|
| recall (%) | CAN | Video | Fusion | CAN | Video | Fusion |
| anger | 74.2 | 74.7 | 76.2 | 70.5 | 69.9 | 67.8 |
| disgust | 74.6 | 75.5 | 77.0 | 71.9 | 72.5 | 69.2 |
| fear | 68.7 | 68.1 | 69.9 | 46.5 | 47.1 | 46.3 |
| joy | 69.1 | 69.9 | 70.1 | 75.6 | 69.6 | 71.4 |
| neutral | 65.7 | 67.3 | 68.9 | 59.3 | 58.9 | 60.8 |
| sadness | 70.8 | 73.3 | 72.6 | 70.1 | 68.4 | 68.0 |
| surprise | 69.2 | 69.4 | 70.6 | 48.5 | 48.5 | 48.3 |
| valence | 64.8 | 66.4 | 68.0 | 56.7 | 56.3 | 58.5 |
| average | 69.6 | 70.6 | 71.7 | 62.4 | 61.4 | 61.3 |