Study of Human Comfort in Autonomous Vehicles Using Wearable Sensors

Haotian Su and Yunyi Jia, *Senior Member, IEEE*

***Abstract*— The rapid development of autonomous vehicles (AVs) has depicted a promising future of a safer and more efficient transportation system. To better induce this revolution, massive efforts have been spent on the technical competence of AVs. However, the human comfort in AVs has been an under-discussed yet important topic to the user acceptance of AVs. For the detection of human comfort, existing studies focus more on the physical influential factors of comfort such as sitting posture, vibration, and noise. With the introduction of AVs, psychological factors also have gained greater influence on human comfort. Despite existing studies of exploring correlations between human comfort and some physiological signals in automated driving contexts, there is few study on how human comfort level in AVs can be detected with these physiological signals. In this paper, we developed effective human comfort study approaches in autonomous vehicles with wearable sensors. We also proposed a machine learning based approach with adaptive feature selection to detect human comfort levels based on the wearable sensing data. The experimental results illustrated the effectiveness of the proposed approaches in studying human comfort in AVs.**

***Index Terms*— Autonomous vehicles, human comfort detection, physiological signal, virtual environment.**

I. Introduction

AUTONOMOUS vehicles (AVs) are at their dawn before a great revolution to the transportation system. The benefits promised by the prevailing of AVs are plenty, including freeing the drivers from drowsy or distracted driving, increased safety, less energy consumption and pollution [1], optimized road capacity [2]. Despite the magnificent effort that the technical experts have dedicated to the enhancement of the AVs' safety and efficiency, the user's acceptance is still a big obstacle on the road of AVs [3]. According to the J.D. Power 2019 Mobility Confidence Index Study [4], the mainstream attitude towards AVs from the customers is still pessimistic. The lack of confidence is pronounced when it comes to comfort. The two questions in the study about the customer's confidence of ride comfort in AVs received 34 and 35 out of 100 points per each. Considering the tremendous works done from the technical aspect and legislation aspect to build up the customers' confidence towards the AVs, the comfort aspect of the AVs seems not receiving enough attention.

A proper platform is needed for carrying out studies on human comfort in AVs. A driving simulator that can provide a realistic virtual environment is an appropriate solution for this type of study. Several studies in the human factors field have proved the feasibility of using a driving simulator instead of an actual vehicle in a human-centered study. Furthermore, driving simulators with high-end and low-cost configurations have been well validated for inducing inner arousal on the participants in these studies. In Bellem's work [5], a full-vehicle scale driving simulator with a motion system was validated for ride comfort study in a comparative experiment between the simulator and actual test drives. Du *et al.* [6] studied driver's trust, preference, anxiety, and mental workload during AV rides with a full-vehicle scale driving simulator. The studies by Peterson employed a simple simulator consisting of a screen, a seat, and a set of steering wheel and pedals to investigate the influence of risk [7] and situational awareness [8] on driver's trust towards a driver's assistance system. These are the successful cases where a driving simulator showed the value in creating a virtual driving environment for human-centered study.

To further study comfort, we should clarify the concept of comfort and determine a way to quantify it. There has not been a standard and universal scientific definition of comfort [5]. In different studies, comfort and discomfort have been interpreted as either two endpoints on a single dimension [9], [10] or as two entirely different constructs that can coexist, each with their own influential factors [11], [12]. While the definition of comfort has been left as an open option for different studies, according to de Looze's work [11], there are three assumptions that have been commonly accepted about comfort: (1) comfort is subjectively defined by each occupant, thus variability between individuals is expected; (2) comfort can be influenced by a wide variety of factors; (3) comfort is influenced by one's reactions to their surroundings and environment.

Before the rapid development of AVs, human comfort studies in vehicles mainly focused on the physical aspects of comfort. These topics include vibration [13], noise [14], thermal comfort [15], air quality [16], etc. Early discussions on human comfort under the AV context focused on some

Manuscript received July 3, 2020; revised December 24, 2020 and June 4, 2021; accepted July 15, 2021. This work was supported in part by the National Science Foundation under Grant CNS-1755771 and Grant IIS-1845779. The Associate Editor for this article was B. F. Ciuffo. *(Corresponding author: Haotian Su.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Review Board of Clemson University under Application No. IRB2017-233.

The authors are with the Department of Automotive Engineering, Clemson University, Greenville, SC 29607 USA (e-mail: haotias@g.clemson.edu; yunyij@clemson.edu).

Digital Object Identifier 10.1109/TITS.2021.3104827

specific control parameters of the vehicle. In [17], a semi-automated-driving scenario was created, and only some simple variations of factors were provided in the scenario to study the relationship between time headway and driver comfort. In [18], one major change in the comfort study of AVs compared to the traditional ones was pointed out. Due to the deprivation of the human driver's control over the vehicle, some psychological factors, e.g., perceived safety, have started to play a bigger role in human comfort in AVs. Therefore, there is a possibility of referring to psychological arousal detection methods in detecting human comfort level in AVs.

In the field of psychological research, multiple studies [19], [20] have examined the effectiveness of using different machine learning or deep learning methods to recognize different types of human emotions with physiological signals. Maaoui's work [21] used two methods, support vector machine (SVM) and Fisher discriminant, to recognize human's emotion of amusement, contentment, disgust, fear, neutral, and sadness with multiple physiological signals, including Blood Volume Pulse (BVP), Electrodermal Activities (EDA), Skin Temperature (SKT). The recognition results for different types of emotions turned out to be excellent, with an accuracy of around 92%. Jang *et al.* [22] and her team attempted to use several different statistical and machine learning methods to examine the differences between several types of emotions with several physiological signals. The study successfully demonstrated the differences between the emotions and found out that discriminant function analysis was the best recognizer of the emotions in their study.

In human comfort detection studies, physiological signals have also been proved to be effective indicators of human comfort. The effectiveness of using various physiological signals, including Electroencephalogram (EEG), EDA, and SKT, for human thermal comfort assessment has been discussed in [23] and [24]. In [25], EEG and EDA were used in the analysis of human comfort in bicycling. In [26]–[28], researchers have studied the correlations between human discomfort and facial expressions and some physiological indicators such as heart rate, pupil dilation, eye blink and EDA in automated driving contexts. However, few work has explored the in-situ detection of human comfort levels in autonomous vehicles using physiological signals. Such a human comfort detection is more challenging than just correlations because it will require much more complex modeling of the relationship between human comfort levels and various physiological signals. Therefore, our research aims to formulate a realistic fully autonomous driving context for the participants and develop models that could perform multiple-level comfort detection based on physiological signals from wearable sensors.

The study was carried out using a high-fidelity driving simulator under the AV context. In this study, we defined human comfort as a single dimension measurement with comfort and extreme discomfort as the two endpoints of a continuum. The definition of human comfort in AVs is further elaborated in Section II. A collection of autonomous vehicle video journeys with synchronized motions were used as the stimuli in the experiment. The data captured in this study included the subjective comfort level and the physiological signals. We designed and produced a pressing button for the participants to press during the ride where a hard press represented a feeling of discomfort about the ride, and not pressing the button meant feeling comfortable about the ride. The physiological signals were collected using a set of wearable sensing devices. After completing the data collection process, we explored the feasibility and effectiveness of using machine learning algorithms to infer the comfort level of the participants with their physiological data.

In summary, the contributions of the paper can be summarized as:

1) Developed and conducted human comfort studies on autonomous vehicles using wearable sensing on a high-fidelity autonomous driving simulation platform.
2) Proposed a support vector machine based approach with adaptive feature selection to detect human comfort levels in AVs using physiological signals from wearable sensors.
3) Conducted a variety of experimental studies to validate and evaluate the proposed comfort detection approaches.

II. Definition of Human Comfort in AVs

Human comfort can be defined in various ways, as several studies have done [9], [11]. The comfort and discomfort caused by vehicle vibrations were interpreted as one continuum with two ends representing very comfortable and very uncomfortable in Oborne's study [9]. While in [11], de Looze modeled the comfort and discomfort when sitting on a chair as two independent constructs, each with its own influential factors.

In this study, we focused on human comfort in AVs. The influential factors that contribute to human comfort in this situation are different from the studies mentioned above. To clarify the scope of comfort we intended to cover, we established a definition of human comfort in AVs.

- Definition: when riding in an AV, human comfort is a feeling of not being unsafe and/or unnatural resulting from the behaviors of the AV itself and the way the AV interacts with the environment.

In this definition, human comfort C in AVs was defined as a single-dimensional construct with one end representing not feeling uncomfortable and the other end representing feeling very uncomfortable. The factors related to human comfort were limited to the behaviors of the vehicle. Only the impacts from the vehicle behaviors to human comfort would be considered in this study.

We defined being comfortable as being free from discomfort feelings because discomfort has been found out to have the dominant effect in human comfort [29]. In [29], Helander studied comfort and discomfort as two independent constructs. While low discomfort ratings were found to coexist with all range of comfort ratings, the comfort rating was found to drop sharply as the discomfort rating increased. This indicated that the factors related to comfort became trivial in the occurrence of discomfort-related factors. Based on this understanding, we believed that identifying discomfort feelings was more important. Therefore, the human comfort in this paper was defined in a single-dimensional way with emphasis on the

Fig. 1. Visualization of the video stimuli. The black lines represent the frames of the monitors.

discomfort feelings, in which feeling comfortable was defined as not feeling discomfort.

In this study, we considered the human comfort changes related to vehicle behaviors only. Elbanhawi *et al.* [18] pointed out that the loss of controllability could expose the passenger of AVs to new influential factors related to discomfort. Compared to traditional human-driven vehicles in which the behaviors can be easily handled, the passenger in AVs would be more likely to feel uncomfortable due to the loss of controllability when experiencing risky or unnatural driving behaviors. The study by Hartwich *et al.* [30] has also suggested that the familiarity of the driving style of the AV could play an important role in a person's preference of an AV. Different opinions towards the behaviors of the vehicle are the major difference between human comfort in a traditional vehicle and an AV. Despite that the factors of human comfort in a traditional vehicle would retain the influence in an AV [21], we focused more on the factors that are unique to AVs and limited the scope the human comfort to comfort only related to the behavioral factors of AVs. The comfort changes resulting from the other factors, e.g. the motion sickness when riding the experimental platform or the room temperature, were not within the range of consideration in this study.

III. Experiment and Data Acquisition

A. Autonomous Driving Stimuli

We created 27 video journeys of simulated autonomous vehicle rides with synchronized motions as the stimuli of the study. A visualization of the video stimuli is shown in Fig. 1. The videos with motions were recorded in a driving simulation software. The duration of each journey was between three to five minutes. Different combinations of road types and driving styles were included in the journeys.

The road types chosen for the study were *city* roads, *highway* roads, and *mountain/rural* roads. The traffic generated by the driving simulation software was in a realistic form with environmental vehicles, road systems, and traffic rules to follow. Three different routes were designed for each type of road. Therefore, there were in total nine different routes in the journeys.

Driving styles were divided into three classes as well, which were named as the *gentle* style, the *normal* style, and the *aggressive* style. Under the *gentle* and *aggressive* style, the simulated vehicle was controlled by the autonomous controller in the simulation software. The vehicle in *normal* style was controlled manually by the experimenter.

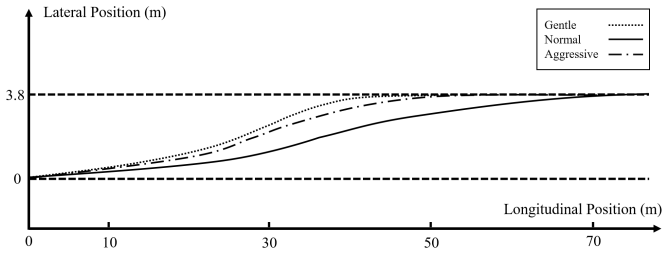


Fig. 2. Trajectories of lane switching maneuvers under three different driving styles. The vehicle was driving at the *highway* TCVs.

We found that the vehicle controlled by the autonomous controller maneuvered with high yaw rate in lateral maneuvers like lane changing and cornering during the creation of stimuli. This led to the unnaturalness in the behaviors of the vehicle controlled by the controller. In order to give a more human-like natural control than what the autonomous controller could provide, we have decided to create the 'normal' driving style by manually driving the agent vehicle in the simulation software. The experimenter acted as the autonomous controller and was required to drive the vehicle with the intermediate configuration that would have been applied to the autonomous controller. Sample trajectories of lane switching maneuvers under different driving styles are displayed in Fig. 2. From the figure, the differences between human-driven style (*normal* style) and automated-controlled styles (*gentle* style and *aggressive* style) could be clearly seen. Besides, we have conducted pre-testing with the experimenter to examine the fidelity of the stimuli. We have received some feedback on the apparent differences between human-controlled and automated-controlled journeys. Based on these facts, the driver's driving style could be clearly distinguished from the other two driving styles controlled by the autonomous controller.

The major differences between the three driving styles were target cruising velocity (TCV), overtaking tendency (OT), and lateral maneuvering quickness (LMQ). The TCV was the cruising velocity of the vehicle when not being blocked by other vehicles and varied for different driving styles and road conditions. The vehicle in the *gentle* style had the lowest OT and never overtook the leading vehicle in the same lane; the vehicle in the *aggressive* would take chances of overtaking on two-lane roads when the rule allowed; the vehicle in *normal*

TABLE I
CONFIGURATION SETUPS OF THE THREE DRIVING STYLES

Configuration	Driving Style		
	gentle	*normal*	*aggressive*
TCV (*highway*)	55 mph	70 mph	85 mph
TCV (*city*)	30 mph	35 mph	40 mph
TCV (*mountain/rural*)	30 mph	35 mph	40 mph
OT	Low	Mid	Hi
Number of Overtaking (*highway*)	2	15	22
LMQ	Hi	Low	Hi

style only took chances of overtaking slower cars on multi-lane roads. The vehicle in *aggressive* style had the highest TCV and OT, and vice versa for the *gentle* style. These features for the *normal* style vehicle were at the intermediate level. The driving styles controlled by the autonomous controller had a high yaw rate when performing lateral maneuvers, while the *normal* style had a lower LMQ. Detailed information about how different driving styles differed from each other is included in Tab. I. The number of overtake maneuvers during highway journeys has been counted and displayed in Tab. I to numerically demonstrate the OTs of different driving styles.

B. Autonomous Driving Simulator

The video stimuli were presented to the participants on a driving simulator with a 3-screen media system, shown in Fig.3. The 6-degree-of-freedom (6-DOF) motions synchronized with the videos, including roll, pitch, yaw, heave, sway, and surge, were also provided to the people riding on the simulator.

The driving simulator used in this study was the Prosimu T6 simulation platform. The platform consists of a moving platform and a static base. A linkage mechanism connects the platform and the base and allows the motors on the static base to generate the 6-DOF motions of the platform. During the experiment, the participant was seated in the racing seat on the moving platform. The seat is equipped with a 4-point harness. The harness was fastened up to guarantee the participant's safety during the experiment as the platform moved. There is an emergency stop button within easy reach of the participants. The platform would return to its neutral position and keep static after the button is pressed down in any unexpected event.

The media system of the simulator consists of a set of three 27-inches monitors. The monitors are aligned horizontally to create an immersive visual experience of wide field-of-view for the participants. Built-in speakers of the monitors and noise-canceling headphones are available options for audio output. Given that the participants in this study needed to wear a physiological data collection device on the head, the built-in speakers were used for audio output in the experiment. In the experiment, the simulator and media system worked collaboratively to deliver the stimuli of simulated autonomous vehicle rides to the participants.

C. Comfort Study Procedures and Protocols

Ten healthy (nine males and one female) and with a mean age of 26.7 (SD 3.68) years old graduate students and faculty members participated in the experiment. All participants had an engineering background and were licensed to drive in the United States.

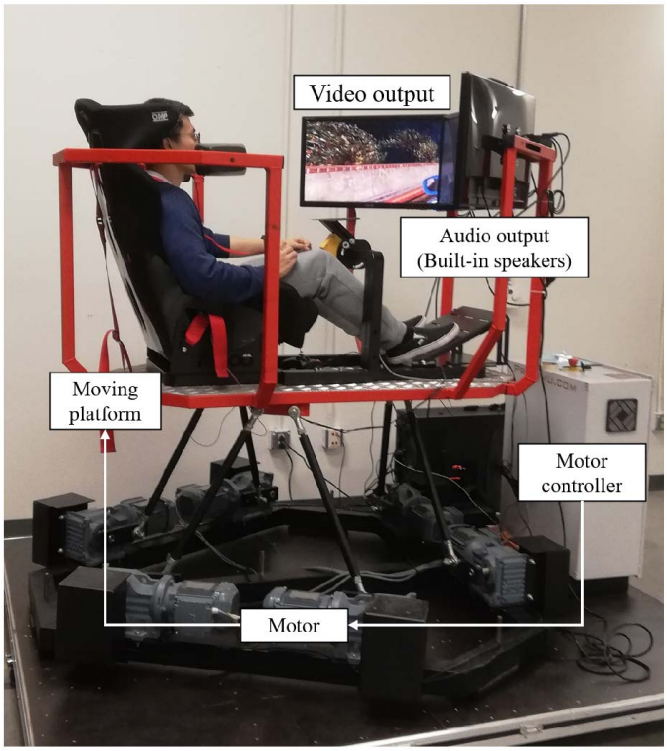


Fig. 3. High-fidelity autonomous driving simulation platform.

Before the experiment, the participants were introduced to the experimental protocols and signed on the consent form for participating in the study. After that, the experimenter should help them wear the physiological measurement devices and assist them to get on the simulator.

Then, the experiment procedures and the tasks for the participants were introduced in detail to them. They were instructed to imagine themselves actually being in a fully autonomous vehicle and experiencing the journey shown in the video. They were also guided to use a pressing button to rate their level of comfort during the virtual ride subjectively. Since the vehicle was assumed to satisfy the SAE Level 5 definition, the participants were not required to perform any takeover action during the experiment. The only task for the participants during the experiment was to provide their comfort levels perceived in real-time based on the behaviors of the vehicle. The data acquisition was initiated after the introduction being addressed.

After the participants responded that they were ready, the experimenter should start to play the videos. Each video lasted for three to five minutes. The participants were asked questions on how they felt after each video played to them. We used a Likert-Scale-based questionnaire from the questionnaire developed in [31] to monitor any sign of the participants suffering from motion sickness. The questionnaire was used for physical condition monitoring purposes only. The results were only used for filtering out data affected by motion

sickness. Their unpleasant feelings of dizzy, queasy, sweaty, and feeling like vomiting would be recorded. If there were no significant unpleasant feelings from the participants, the next video would be played when they were ready; otherwise, the participants could take a break until they felt physically ready for the next video. The participants had the right to suspend or withdraw from the experiment at any time.

The whole experiment was divided into three separate sessions to control the tiredness of the participants during the process. The typical duration of one session in the experiment was 75 minutes. One participant could get scheduled with only one session at most per day. Nine journeys were included in one session. The participants only needed to fill out the consent form during their first session. The instructions were repeated in a brief version in the next two sessions of the experiment.

The design of how different road types and driving styles were presented to the participants was optimized to help the participants better adapt to the experimental environment. The journeys within one session were all from the same type of road. We expected the participants to maintain a consistent criterion for feeling comfortable or uncomfortable about the maneuvers occurring within a specific type of road when experiencing only one type of road in one session.

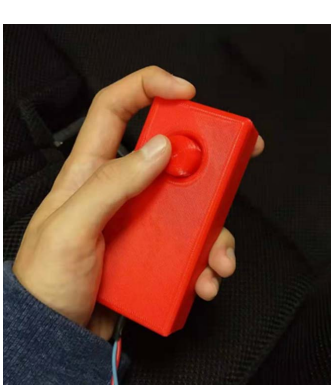
In the first session, the *highway* journeys were experienced by the participants, and the three driving styles were arranged in the order of the *gentle* style, the *aggressive* style, and then the *normal* style. The participants were in the process of getting familiar with the virtual experience in the first session. A simpler scenario should help the participant adapt to the experience. Therefore, the *highway* road type, which was the simplest within the three road types, was selected for the first session. The three journeys in the *gentle* style were presented in the beginning as an adaptation to the experience for the participants. The journeys in the *normal* style followed the *gentle* style, and the *aggressive* style was the last style presented. We designed the order that gradually moved from the least to the most aggressive style to help the participants gradually obtain a sense of the differences between the driving styles and establish a criterion for perceiving and expressing their comfort levels.

For the next two sessions, we assumed that the participants had completed the adaptation process and established a criterion for comfort and discomfort, and the orders of road types and driving styles were determined randomly. The rules were still obeyed that only one road type was included within one session and the journeys under the same driving style were presented in a row. The orders of the road types and driving styles were the same for all participants for the ease of data processing. The arrangement of how different road types and driving styles were presented to the participants is demonstrated in Tab. II. However, the orders of scenarios in each session for a participant were different. Furthermore, since each participant participated in this experiment only once, the scenarios they saw were always new to them. The potential risk of learning effects and bias would only occur if a participant participated in the same experiment multiple times. However, we did not have such a case in our experiment because it would violate the rule of keeping the participants innocent of the stimuli.

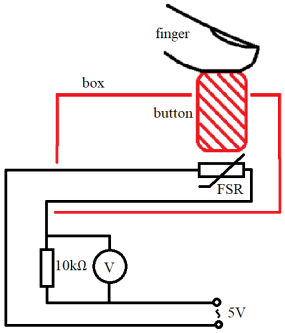
TABLE II

ARRANGEMENT OF JOURNEYS IN DIFFERENT ROAD TYPES AND DRIVING STYLES IN THE THREE EXPERIMENTAL SESSIONS

	# Session		
	1	2	3
Road Type	*highway*	*city*	*mountain/rural*
1^{st} Driving Style	*gentle*	*normal*	*aggressive*
2^{nd} Driving Style	*normal*	*aggressive*	*gentle*
3^{rd} Driving Style	*aggressive*	*gentle*	*normal*



(a) Picture of the button box.



(b) Structure diagram of the button box.

Fig. 4. Physical button box for pressing force collection and its structure diagram.

D. Subjective Comfort Level Acquisition

Questionnaires have been proved to be an effective approach for ground truth comfort level collection in different studies [32], [33]. However, the participants would have been interrupted from the AV journey experience when answering the questions. Also, it is impossible to collect real-time data with questionnaires. In this study, we hoped that the participants could maintain an undisturbed immersive experience while providing the real-time comfort level report. With the inspiration from [34], a button box was designed and produced for the experiment to collect real-time comfort from the participants, as is shown in Fig. 4.

The box was 3D printed about the size of a smartphone, 106 mm length, 51 mm width, and 23 mm height. All edges were rounded to fillets of a 2 mm radius for better holding comfort. A button with a radius of 5 mm was located within the natural reach of a person's thumb. The top surface of the button was curved to fit the shape of a thumb. This design would contribute to the comfort of usage as well.

A force-sensitive-resistance (FSR) was placed between the button and the box. The resistance of the FSR would drop as the pressure being applied on it increased. We used a National Instruments USB-6003 Data Acquisition (DAQ) device to capture the resistance change on the FSR to infer the pressing force applied by the participants. We selected the sampling rate as 15 Hz with the consideration of how the comfort

levels of the participants were stimulated. The comfort changes considered in this study were expected to result from the behaviors of the AV. Since the behaviors of the AV should not change at a very high rate during the journey, the sampling rate was selected at 15-Hz to save the hardware resource.

The participants were instructed to press the button whenever they felt uncomfortable about the behaviors of the vehicle during the ride. They should press and hold the button until the uncomfortable feeling disappeared. The more uncomfortable the participants felt, the harder should they press the button. They were also asked to keep clear of the button when there was no uncomfortable feeling. Because the FSR was very sensitive and unintentional touches on the button could confuse the measurement. We used the pressing force from the participants as the measurement of the participants' real-time subjective comfort level, where a higher pressing force value represented a higher level of discomfort and vice versa.

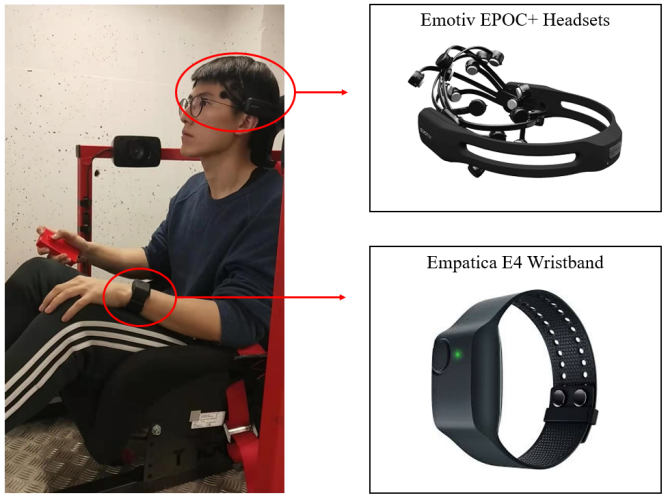


Fig. 5. Wearable sensing devices used in the study: Emotive EPOC+ headsets and Empatica E4 wristband.

E. Data Acquisition With Wearable Sensing Devices

Two types of wearable sensing devices were used in this study, as shown in Fig. 5, the Empatica E4 wristband and the Emotiv EPOC+ headsets. With these devices, a wide range of physiological signals was captured in the experiment. Given that the EDA, SKT, and EEG have been used in different studies [23]–[25] for comfort detection purpose, and that the EDA, BVP, and SKT have been employed in emotion detection studies [21], [22], we selected EEG, EDA, BVP, and SKT for the human comfort detection in this study. These signals have been validated to be effective in multiple studies and thus would be reliable for the usage in this study.

EEG provides us the insight into the brain activities of a person [35]. It is measured by applying multiple non-invasive electrodes on a person's scalp. The electric potential values measured from the skin at different locations on a person's scalp reflect the activities taking part in the corresponding parts of the brain. The EEG data in this study were captured using the Emotiv EPOC+ headsets. During the measurement process, the 16 fabric feelers of the headsets were saturated with saline to enhance the conductivity between the feelers and the skin. The EmotivPro software worked with the headsets and output the measurement results, including raw EEG data, frequency domain analysis of raw EEG data, and performance metrics that measured the person's mental state. The frequency-domain analysis of EEG data is usually carried out based on the band power of five different frequency bands ($\delta : 1-3$Hz, $\theta : 4-7$Hz, $\alpha : 8-13$Hz, $\beta : 14-30$Hz, $\gamma : 31-50$Hz) of the signals from different measurement channels. A short-time Fourier transform can be performed on a given sequence of EEG measurements first. Then the powers of the five frequency bands can be manipulated to get the features for further usages like emotion recognition [35]. The performance metrics signals provided by the EmotivPro software include stress, engagement, interest, focus, excitement, and relaxation. These signals are directly outputted from the software, and each represents the arousal of a specific type. In [36], the researchers successfully used the performance metrics signals to differentiate the brain activities between the rest and active states. Aspinall and his team [37] used the Emotiv EPOC headsets for an outdoor experiment and captured a systematic difference in the performance metrics signals as the participants walked in different urban areas. We used the excitement (EXC) from the performance metrics for the feature extraction, which was believed to be related to positive feelings of arousal. The EXC had a 0.1-Hz sampling rate.

EDA is a physiological signal that can be easily measured from the skin. This signal characterizes the change in the electric conductance of the skin due to the activity of the autonomic nervous system. The arousal of the sympathetic autonomic nervous system activity can increase sweat gland activity, which leads to greater skin conductance. Thus, the EDA signal can be used as an indication of a person's psychological or physiological arousal in response to an external stimulus [38]. The EDA signal was measured with the Empatica E4 wristband. Although the wrist area is not the top-recommended area for EDA measurement [39], we selected the wrist measurement solution of Empatica E4 because the E4 wristband provided an integrated measurement of physiological signals. Using E4 saved extra devices needed for BVP and SKT measurement. The wireless measurement also ensured the least obtrusiveness to the participants during the experiment. The wrist has been recognized as the best alternative area for EDA measurement when neither fingers nor the feet are available [39]. With these considerations, the wrist measurement provided by Empatica E4 was selected for the EDA measurement. For the measurement, two AgCl plated electrodes were located inside the strap. The electrodes touched the skin of the inside of the wrist and lined up under the middle and ring fingers. The data were measured from the non-dominant hand of the participants and had a 4-Hz sampling rate.

BVP measures the heart rate based on the volume of blood passing through the tissue at each beat of the heart. It is measured with the photoplethysmography (PPG) sensor

embedded in the Empatica E4 wristband. The PPG sensor sends out an infrared light onto the surface of the skin on the wrist. The light transmits through the tissue and gets reflected back to the sensor. Because the hemoglobin selectively absorbs the light in the red blood cell, the amount of light reflected will be proportional to the volume of the blood flowing under the tissue. The pulse in the measurement of the reflection light from the PPG sensor is then processed into BVP data. During the experiment, the measurement of BVP had a 64-Hz sampling rate.

SKT measures the thermal changes on the skin. The SKT variations mainly result from localized changes in blood flow caused by vascular resistance or arterial blood pressure. Local vascular resistance is modulated by smooth muscle tone, which is mediated by the sympathetic nervous system. The mechanism of arterial blood pressure variation can be described by a complicated model of cardiovascular regulation by the autonomic nervous system. The SKT variation reflects autonomic nervous system activities and is another indicator of a person's psychological state [22]. During the experiment, the SKT had a 4-Hz sampling rate.

F. Preprocessing of Data

Before feature extraction, several preprocessing tasks were completed upon the raw data we had. For the subjective comfort level data, because the participants had different physical condition, the pressing force collected were standardized to eliminate the individual differences. The Z-Score methodology was used for the process. Supposing that the comfort level data from a participant being denoted as a set $\{C_p\}$, a new set $\{C_{p\_new}\}$ including the original $\{C_p\}$ and a new set $\{-C_p\}$ consisting of the opposite numbers of all elements in the original set. The standardization results can be calculated through the following formula:

$$z_p = (c_p - \mu)/\sigma \tag{1}$$

where z_p is the standardized comfort level which is a dimensionless number, c_p is the raw comfort level from the original set $\{C_p\}$, μ is the mean value of the set $\{C_{p\_new}\}$ which equals to zero, and σ is the standard deviation of the set $\{C_{p\_new}\}$. The labeling process of the dataset was based on the standardized comfort level data.

The EDA signal consists of the long-term slow-changing component (skin conductance level, SCL, also known as the tonic component) and the short-term fast-changing component (skin conductance responses, SCRs, also known as the phasic component) [40]. Both components were found to be well correlated with emotional changes [41], [42].

To extract these two components, a Matlab-based software, Ledalab, was used to preprocess the EDA signal. This software employed continuous decomposition analysis (CDA) [40]. The first step of CDA was to perform deconvolution on EDA data. The nerve activities that cause EDA can be considered a driver, consisting of a series of impulses that trigger a specific impulse response, i.e., SCRs. Suppose an impulse response function (IRF) that describes the course of the impulse response over time. The result of this process can be represented by the convolution of the driver with the IRF:

$$SCR = Driver_{phasic} * IRF \tag{2}$$

Since the EDA signal consists of both tonic and phasic component, we have:

$$EDA = SCL + SCR = SCL + Driver_{phasic} * IRF \tag{3}$$

The tonic component can be equally represented as the convolution of a tonic driver function and the same IRF. The EDA data can be represented by:

$$EDA = (Driver_{tonic} + Driver_{phasic}) * IRF \tag{4}$$

Deconvolution is the reverse process of convolution. The deconvolution of the EDA data results in a composite of both tonic and phasic driver functions. After either of the two functions has been estimated, the other one can be calculated by subtracting the known driver from the total driver as well:

$$\frac{EDA}{IRF} = Driver = Driver_{tonic} + Driver_{phasic} \tag{5}$$

With the driver estimated, the corresponding data can be calculated by the convolution of the driver and the IRF. The raw EDA signal can then be decomposed into two compositions. For the data process in this study, the EDA, SCL, SCR signals were all involved in the feature extraction process.

G. Label Creation

To perform label creation, the subjective comfort level data were cut into samples with a time duration t_s. The label of each segmentation was based on the standardized comfort level data within the segmentation. The creation of labels followed the two rules:

- For binary labels: the segmentation should be labeled as 'discomfort' whenever a comfort level peak greater than 0.1 was spotted within a segmentation; otherwise, it should be labeled as 'comfort'.
- For 3-level labels: for samples previously labeled as 'discomfort', if the mean comfort level was no greater than 0.5, then it should be labeled as 'low-discomfort'; if the mean comfort level was greater than 0.5 and no greater than 3, then it should be labeled as 'mid-discomfort'; if the mean comfort level was greater than 3, then it should be labeled as 'hi-discomfort'; no changes to samples of 'comfort'.

The cut-off value between 'comfort' and 'discomfort' samples was determined based on the measurement noise of the DAQ device. When not being pressed, the measurements of the noises from the DAQ device were typically small numbers under 0.03 after standardization. Furthermore, any pressing feedback that indicated a feeling of discomfort should result in a measurement of greater than 0.15 after standardization. Based on this observation of the data, we concluded that setting the cut-off value at 0.1 could effectively distinguish 'discomfort' samples from the random noises. Therefore, 0.1 was selected as the cut-off value between 'comfort' and 'discomfort' samples.

The cut-off values between the three levels of discomfort samples were determined based on the standardization process of the data. The Z-Score value of a data point shows how far the point is from the average value. Based on our standardization process, the value should demonstrate how far a point was from being a 'comfort' sample. If the Z-Score of a 'discomfort' sample was under 0.5, we considered it as a 'low-discomfort' sample that was close to 'comfort' samples; if the Z-Score of a 'discomfort' sample was between 0.5 and 3, it meant that the comfort level was at 'mid-discomfort' level which was far from being comfortable but not extremely uncomfortable; for the other samples with Z-Score above 3, the comfort levels represented in these samples were extremely far from being comfortable, and thus these samples were considered as 'hi-discomfort' samples. In summary, the Z-Score provided us a standardized way to understand and determine different levels of discomfort, and we chose these cut-off values based on our knowledge of the standardized comfort levels.

H. Feature Extraction

With a wide range of signals available, the next step was to perform feature extraction on the data to prepare the material for the training process. To perform feature extraction, the physiological signal measurements were first synchronized with the comfort level data and cut into samples with the same time duration t_s as the label creation process.

Normalization is also an important technique for minimizing the individual differences in their physiological responses when exposed to the same stimuli. The collected physiological signals were normalized using the average value of the same type of signal collected within all samples labeled as 'comfort' [19]. The normalized physiological signals were calculated using the following formulas:

$$\bar{x}_{com} = \frac{\sum x_{com}}{n_{com}} \tag{6}$$

$$x_{norm} = \frac{x_{raw} - \bar{x}_{com}}{\bar{x}_{com}} \tag{7}$$

where x_{norm} is the normalized signal, x_{raw} is the raw signal, $\bar{x_{com}}$ is the average value of the signal within all samples labeled as 'comfort', x_{com} is the signal within samples labeled as 'comfort', and n_{com} is the number of samples labeled as 'comfort'.

Based on the features proposed in Picard's work [43], several additional statistical features were also considered in the feature extraction process. For an acquired signal x, the features directly taken from Picard's work were the mean value μ_x, the standard deviation σ_x, the mean of the absolute values of the first differences δ_x, and the mean of the absolute values of the second differences γ_x. Other than these, our additional statistical features were the maximal value max_x, the minimal value min_x, the odds of the minimal value over the maximal value $ratio_x$ [44], and the root mean square value M_x. All eight statistical features mentioned above were calculated within each of the sample for each type of physiological signal and its normalized form. These features were calculated with the following formulas:

$$\mu_x = \frac{1}{T}\sum_{t=1}^{T} X(t) = \bar{X}(t) \tag{8}$$

$$\sigma_x = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(X(t) - \mu_x)^2} \tag{9}$$

$$\delta_x = \frac{1}{T-1}\sum_{t=1}^{T-1} |X(t+1) - X(t)| \tag{10}$$

$$\gamma_x = \frac{1}{T-2}\sum_{t=1}^{T-2} |X(t+2) - X(t)| \tag{11}$$

$$max_x = max(x) \tag{12}$$

$$min_x = min(x) \tag{13}$$

$$ratio_x = \frac{min(x)}{max(x)} \tag{14}$$

$$M_x = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(X(t))^2} \tag{15}$$

where $X(t)$ is the signal x at sampling number t, and T is the total number of samples.

Given that the features were extracted from all six sources of raw physiological signals, including the EXC, EDA, SCL, SCR, BVP, and SKT, and all these signals had their normalized form, a total of 96 features was extracted. The sample length t_s for the comfort level data and the physiological signals were selected based on two considerations. A too-short time window would not be able to capture some discomfort feelings, while a too-long time window would introduce noises to the labeling and feature extraction. They would both influence prediction accuracy. To balance the two concerns, we selected 2 seconds as the length of the samples, and there were 3,032 samples for each type of signal per participant.

IV. Comfort Level Detection Methodology

A. Support Vector Machine

SVM is a non-linear model that has been widely applied in various fields. The advantage of SVM is that it uses only the support vectors for finding the optimal hyperplane that maximizes the margin. The use of support vectors reduces the complexity of finding the hyperplane function and increases the generalization of the classifier when the size of the dataset is not large. SVM can also handle non-linear datasets well with the use of the kernel trick.

In different studies, SVM has been proved to be an effective machine learning model to detect comfort [24] and emotions [20], [22]. The ability to handle high-dimensional features is an important advantage of SVM. Besides SVM, some other algorithms have also been tested in [22] and [24]. The algorithms in common between these two studies include Linear Discriminant Analysis (LDA), Naive Bayes (NB), Classification And Regression Tree (CART), and SVM. When applied to different scenarios, the performance of LDA and NB varied a lot, both showing good performance in [22] while performing poorly in [24]. For CART that outperformed SVM

in both studies, the ability to handle high dimensional features is a weakness compared to SVM. Because the effectiveness of SVM has been proved to be consistent in related studies, and SVM is more suitable for processing high dimensional features, which could be up to 96 dimensions in this study, SVM was selected to develop the comfort detection model.

In this study, we used a binary SVM classifier to solve the problem. The kernel function selected for the classifier was the radial-based function (RBF). The input of the classifier was the sample's physiological features, and the output of the classifier was the predicted label of the comfort level of the sample. In the classification process, the decision function is given by:

$$f(x) = sgn(\sum_{i=1}^{n} y_i \alpha_i K(x_i, x) + \rho) \tag{16}$$

where $f(x)$ is the classification results of the sample input x, n is the total number of support vectors, y_i is the label of the i^{th} support vector, α_i is the Lagrange multiplier of a dual optimization problem, x_i is the i^{th} support vector, ρ is the interception, $K(x_i, x)$ is the kernel function, which is RBF in our case:

$$K(x_i, x) = exp(-\Gamma \|x_i - x\|^2) \tag{17}$$

where Γ is the inverse of the standard deviation of the RBF core, which is an important hyperparameter of an SVM classifier.

When processing the training data, we noticed that the dataset was imbalanced. For most of the participants, the proportion of samples labeled as 'discomfort' took up around 20% to 30% of the entire dataset. The number of 'comfort' samples would be two to four times the number of 'discomfort' samples in the training set if no countermeasure were taken. The performance of the classifier would be negatively affected if it was trained with the imbalanced dataset. We employed the under-sampling method to construct a balanced dataset. Within the dataset of a specific participant, an equal amount of 'comfort' samples were randomly selected based on the number of 'discomfort' samples. The 'discomfort' samples and the selected 'comfort' samples were then combined into a new balanced dataset. The data processing for the study was based on the new dataset for each participant.

B. Adaptive Feature Selection

Among the 96 features extracted from the physiological signals, not all of them were valuable for the training of the classifier. Some features might not contribute to or even confuse the training process. Therefore, a feature selection algorithm should be applied to the training data to improve the training results.

We proposed an adaptive feature selection (AFS) methodology using SVM Recursive Feature Elimination (SVM RFE) and Bayesian optimization algorithms. The overall process of the AFS algorithm can be simplified to a flow chart in Fig. 6.

RFE algorithm is a feature ranking algorithm [44]. The process of RFE can be sorted into three steps:

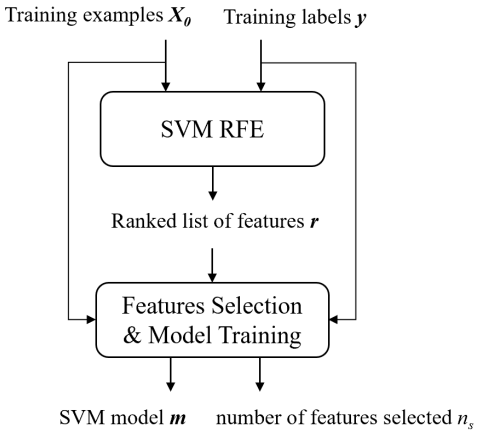


Fig. 6. The working process flow chart of the AFS algorithm.

Algorithm 1 Algorithm for SVM RFE

Input: Training examples $\mathbf{X_0} = [\mathbf{x_1}, \mathbf{x_2}, \ldots \mathbf{x_m}]^T$, class labels $\mathbf{y} = [y_1, y_2, \ldots y_m]^T$.
Output: Ranked list of features $\mathbf{r} = [f_1, f_2, \ldots f_n]^T$.
the indices of features not eliminated $\mathbf{s} \leftarrow [1, 2, \ldots n]^T$, the ranked list $\mathbf{r} \leftarrow [\ ]$.
while s not empty **do**
  $\mathbf{X} \leftarrow \mathbf{X_0}(:, \mathbf{s})$.
  train an SVM classifier **model** $\leftarrow SVM\_train(\mathbf{X}, \mathbf{y})$.
  compute $\mathbf{w} = \sum_k \alpha_k y_k \mathbf{x_k}$.
  compute $c_i = w_i^2$, for all i.
  $f \leftarrow argmin(\mathbf{c})$.
  $\mathbf{r} \leftarrow [\mathbf{s}(f), \mathbf{r}]^T$.
  $\mathbf{s} \leftarrow \mathbf{s}(1 : f-1, f+1 : end)$.
end while
return r.

1) Train the classifier (optimize the weights w_i with respect to cost function J).
2) Compute the ranking criterion for all features (can be w_i^2).
3) Remove the feature with the smallest ranking criterion.

SVM RFE algorithm employs a linear SVM as the classifier in the training process [44]. The rank criterion is calculated with the following formulas:

$$\mathbf{w} = \sum_k \alpha_k y_k \mathbf{x_k} \tag{18}$$

$$c_i = w_i^2 \tag{19}$$

where $\mathbf{w}$ is the weight vector, α_k is the k^{th} Lagrange multiplier, y_k is the label of the k^{th} support vector, $\mathbf{x_k}$ is the k^{th} support vector, c_i is the criterion of the i^{th} feature. The feature with the lowest criterion will be eliminated for the current iteration and be put into the ranked list. After all features have been added to the ranked list, the algorithm stops and returns the list. The algorithm of the SVM RFE is described in Algorithm 1.

With the ranked list where all features are ranked according to their ranking criterion from the highest to the lowest, we will train and tune the classifier using different features.

Algorithm 2 Algorithm for Feature Selection and Model Training

Input: Training examples $\mathbf{X_0} = [\mathbf{x_1}, \mathbf{x_2}, \dots \mathbf{x_m}]^T$, class labels $\mathbf{y} = [y_1, y_2, \dots y_m]^T$, ranked list of features $\mathbf{r} = [f_1, f_2, \dots f_n]^T$.

Output: SVM model trained with optimized number of features $\mathbf{m}$, optimized number of features n_s.

$i \leftarrow 0$, $l_{prev} \leftarrow 0$.
while $i < length(\mathbf{r})$ **do**
  $\mathbf{X} \leftarrow \mathbf{X_0}(:, \mathbf{r}(1:i))$.
  create a K-Fold partition c based on the training set $\mathbf{X}$.
  train an SVM classifier, use Bayesian optimization to tune the classifier based on cross-validation loss **model** $\leftarrow SVM\_train(\mathbf{X}, \mathbf{y}, BayesOpt, KFold = c)$.
  collect training loss of **model**, $l \leftarrow$ **model**.$loss$.
  if $l < l_{prev}$ **then**
    $l_{prev} \leftarrow l$.
    $\mathbf{m} \leftarrow$ **model**.
    $n_s \leftarrow i$.
  end if
end while
return $\mathbf{m}$, n_s.

The kernel function in the SVM used in the study was RBF kernel. Two hyperparameters, C and Γ, were to be tuned in an RBF kernel to achieve the optimal classification result [45]. The C parameter determines the trade-off between minimizing the training error and reducing the complexity of the model. The Γ parameter has been mentioned in (17) from the previous section. It affects the mapping from the input space to the high-dimensional feature space.

Bayesian optimization method was employed in finding the optimal hyperparameters for the classifier. In the Bayesian optimization process [46], the cost function is treated as a random function, and a prior is placed over it. The prior gets updated to form a posterior distribution as the optimization process goes on. The posterior is used for the determination of the next query point in turn. In our case, the prior was based on the 2D plane of the two hyperparameters, C and Γ. The cost of the optimization was the cross-validation loss of the training process. A lower cross-validation loss could help us better approach the parameters that provide the best generalization performance for the classifier.

The training process shall repeat for 96 iterations for each participant, as there are 96 available features in total. The first k features in the ranked list would be used in the k^{th} iteration of the process. After the process was finished, the algorithm would return a list of the cross-validation loss values acquired from each iteration. The number of features used in the iteration with the lowest loss value would be selected as the optimized number of features. The algorithm of feature selection and SVM model training is described in Algorithm 2.

The working process of the AFS algorithm was the combination of Algorithm 1 and 2. It started with ranking the features using the SVM RFE. And then, the algorithm iterated 96 times to train an SVM and optimize the parameters using Bayesian optimization method. One more feature from the ranked list was added to the feature set used for training in the previous iteration during each iteration. After the iterations were done, the number of features that achieved the lowest training loss would be returned as the optimal result.

V. Experimental Results

We carried out the experiment on 10 participants and collected the comfort data and physiological signals required for the comfort level detection. During the experiment, no participant reported any sign of motion sickness with the questionnaire we used. In this study, we performed the AFS algorithm on the data collected from all 10 participants. The ranked lists of features from the participants were all unique to each other.

For all the participants, different numbers of features used in the SVM were tested. The options were 5- features, 10-features, 20-features, 96-features, and the number given by the AFS algorithm.

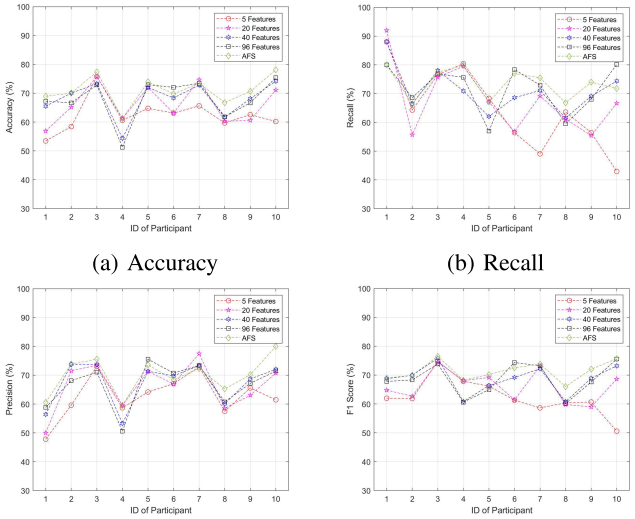
Different formations of the dataset were also tested in this study. The 'comfort' vs. 'discomfort' dataset was included in the first place. Other than that, we also explored the possibility of performing human comfort detection on multiple levels. Since SVM is a binary classifier, the solution to the realization of multi-class classification is to train multiple SVM models to work collaboratively in a multi-class classification task. There are several ways to decompose the multi-class problem into a set of binary classification problems [47], e.g., One-Against-All method, One-Against-One method, Directed Acyclic Graph SVM. In this study, we focused more on the discomfort perceived by the participants and defined three levels of discomfort. The multi-class classifier should be designed to distinguish between different levels of discomfort. Therefore, another three datasets, the 'low-discomfort' vs. 'mid- & hi-discomfort', 'mid-discomfort' vs. 'low- & hi-discomfort', and the 'hi-discomfort' vs. 'low- & mid-discomfort' datasets, were created and tested as well.

All the datasets used in training were balanced using the under-sampling method. The distribution of training samples and testing samples was 70% for training and 30% for testing. The datasets underwent 10-fold cross-validation during the training process.

A. Results of Comfort Level Detection

Fig. 7 gives us the testing accuracy, recall, precision, and F1-Score of comfort level detection when the classifier was trained with 5-features, 10-features, 20-features, all 96-features, and features selected by AFS algorithm. The circles in the figure represent the results with the classifiers trained with 5-features, the stars stand for the results from classifiers trained with 20-features, the hexagram stand for the results from the classifiers trained with 40-features, the squares represent the results from the classifiers trained with 96-features, the diamonds represent the results from the classifiers trained with features selected with the AFS algorithm.

Fig. 8 shows the confusion matrices achieved across all participants with the classifiers trained with different number of features. We can see that the testing datasets were well



(a) Accuracy (b) Recall

(c) Precision (d) F1 Score

Fig. 7. Classifier testing results of all the participants when trained with different no. of features. The circles represent the results obtained with classifiers trained with 5 features; the stars represents the results with 20 features; the hexagrams stand for the results with 40 features; the squares are for 96 features; and the diamonds stand for the results obtained with features selected by AFS.

(a) 5 Features (b) 20 Features

(c) 40 Features (d) 96 Features

(e) AFS

Fig. 8. Confusion matrices based on the testing results across all participants when using classifiers trained with different no. of features.

balanced in the number of 'comfort' and 'discomfort' samples. The performances of the classifiers in classifying samples with different labels were similar. Overall, the under-sampling method helped with balancing the datasets and the performance of the classifiers was well balanced for different classes.

For all the participants, the personal best accuracy of comfort level detection was above 60% considering all available feature sets. The best detection accuracy for seven participants was 70% and greater. Among all the combinations of participants and features, the best accuracy of comfort level detection was 78.12%, achieved from Participant 10 when the AFS was applied. For eight participants, the features selected with the AFS algorithm achieved the highest detection accuracy among all options for the features set. By using the AFS algorithm, the detection of comfort level achieved over 70% accuracy with six participants. For other feature sets that were chosen empirically, the numbers of participants who achieved 70% accuracy or higher were one person for 5-features, four persons for 20-features, five persons for 20-features, and five persons for 96-features.

Tab.III shows the average testing accuracy of the classifier when trained with 5-features, 10-features, 20-features, all 96-features, and features selected with AFS algorithm.

Among all the options of feature combinations tested in the study, the features selected with the AFS algorithm achieved the highest average accuracy across participants. The average accuracy was 71.88% for the features selected with AFS. Besides, we can also see that the standard deviation values of the accuracy across different participants were all above 5% for all the feature selection options.

B. Results of Comfort Level Detection in Three Levels

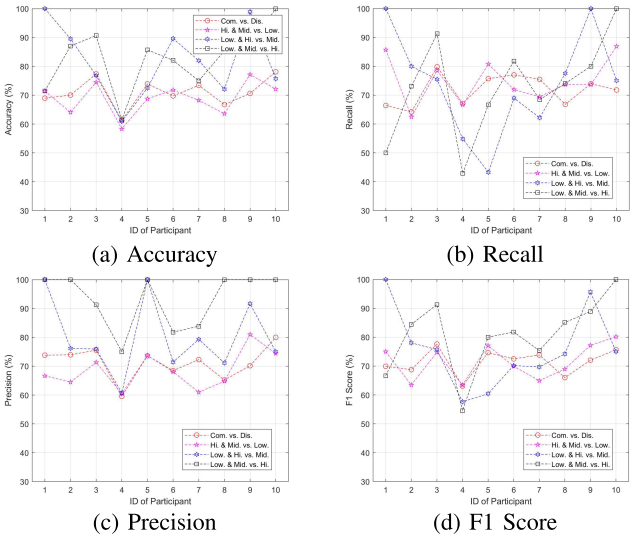
Fig. 9 displays the accuracy, recall, precision, and F1-Score of the classifier trained with different datasets based on the features selected with the AFS algorithm. The circles in the figure represent the results with the classifiers trained with the 'comfort' vs. 'discomfort' datasets; the stars stand for the results from classifiers trained with the 'low-discomfort' vs. 'mid- & hi-discomfort' datasets; the hexagrams stand for the results from the classifiers trained with the 'mid-discomfort' vs. low- & hi-discomfort' datasets; the squares represent the results from the classifiers trained with the 'hi-discomfort' vs. 'low- & mid-discomfort' datasets. The names of these classifiers were abbreviated to 'dis' classifiers, 'low' classifiers, 'mid' classifiers, and 'hi' classifiers in the following.

Fig. 10 shows the confusion matrices achieved across all participants with the classifiers trained with different datasets. The label distribution was less balanced than the sample distribution of the 'comfort' and 'discomfort' samples in the entire dataset. We can see that the performance of the 'low' classifiers was well balanced. For the 'mid' and 'hi' classifiers, although the overall accuracy values were better, the performances of the classifiers were less balanced compared to the 'low' classifiers.

TABLE III

STATISTICS OF THE TESTING ACCURACY, RECALL, PRECISION, AND F1 SCORE OF THE CLASSIFIER WHEN TRAINED WITH 5-FEATURES, 20-FEATURES, 40-FEATURES, 96-FEATURES, AND FEATURES SELECTED WITH AFS

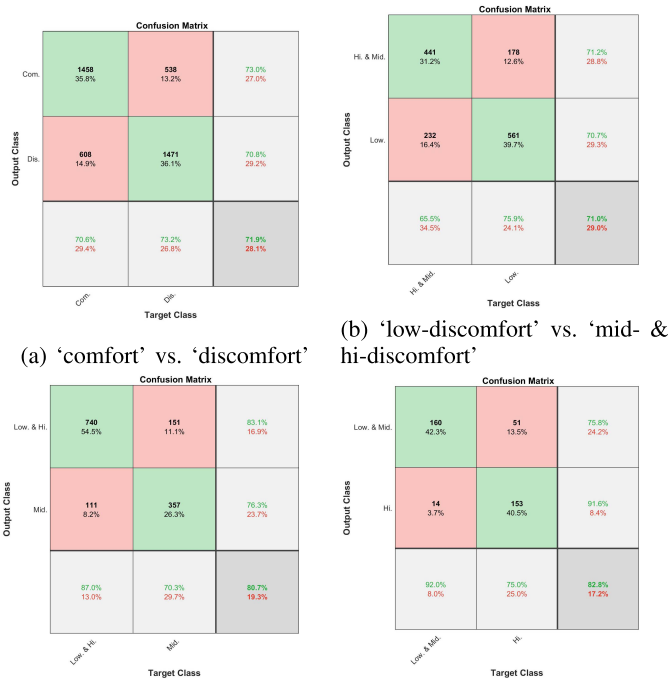
	# Features				
	5	20	40	96	AFS
Avg. Accuracy	64.29	66.85	69.01	69.13	71.88
Min. Accuracy	53.45	56.90	54.38	51.25	61.25
Max. Accuracy	75.77	75.77	74.16	75.38	78.12
Accuracy SD	5.83	6.74	6.16	7.23	5.01
Avg. Recall	62.06	64.11	69.75	70.84	73.59
Min. Recall	42.95	55.41	61.54	59.51	66.43
Max. Recall	80.42	79.53	78.04	80.13	79.82
Recall SD	14.14	11.81	7.82	8.28	5.36
Avg. Precision	64.64	67.15	68.84	68.00	71.22
Min. Precision	47.83	50.00	53.26	50.56	59.55
Max. Precision	73.05	77.46	73.81	75.52	80.00
Precision SD	7.57	8.37	7.72	7.75	5.75
Avg. F1 Score	63.32	65.60	69.29	69.39	72.39
Min. F1 Score	50.57	58.98	60.80	60.12	63.10
Max. F1 Score	74.76	74.53	75.84	75.53	77.63
F1 Score SD	6.32	5.43	4.89	5.58	4.50



(a) Accuracy (b) Recall (c) Precision (d) F1 Score

Fig. 9. Classifier testing results of all the participants when trained with different datasets. The circles represent the results obtained with 'dis' classifiers; the stars represents the results with 'low' classifiers; the hexagrams stand for the results with 'mid' classifiers; and the squares stand for the results obtained with 'hi' classifiers.

Compared to the 'dis' classifiers, an overall increase of test accuracy from the other classifiers can be observed from the figures. More specifically, the 'low' classifiers of seven participants only achieved a slightly worse accuracy compared to the 'dis' classifiers; the 'mid' classifiers of six participants outperformed the 'dis' classifiers; all participants ended up with better accuracy with the 'hi' classifiers compared to the 'dis' classifiers. Five participants achieved accuracy over 80% with the 'mid' classifiers, and seven participants achieved accuracy over 80% with the 'hi' classifiers.



(a) 'comfort' vs. 'discomfort' (b) 'low-discomfort' vs. 'mid- & hi-discomfort' (c) 'mid-discomfort' vs. low- & hi-discomfort' (d) 'hi-discomfort' vs. 'low- & mid-discomfort'

Fig. 10. Confusion matrices based on the testing results across all participants when using classifiers trained within different datasets.

TABLE IV

STATISTICS OF THE TESTING ACCURACY, RECALL, PRECISION, AND F1 SCORE OF THE 'DIS', 'LOW', 'MID', AND 'HI' CLASSIFIERS

	Classifier			
	'dis'	'low'	'mid'	'hi'
Avg. Accuracy	71.88	70.96	80.72	82.80
Min. Accuracy	61.25	58.33	60.94	61.54
Max. Accuracy	78.12	77.17	100.00	100.00
Accuracy SD	5.01	5.66	12.58	10.82
Avg. Recall	73.59	75.91	70.28	75.00
Min. Recall	66.43	62.50	43.28	42.86
Max. Recall	79.82	86.92	100.00	100.00
Recall SD	5.36	7.97	17.89	17.29
Avg. Precision	71.22	70.74	76.28	91.62
Min. Precision	59.55	60.47	60.71	75.00
Max. Precision	80.00	80.99	100.00	100.00
Precision SD	5.75	6.50	12.98	9.60
Avg. F1 Score	72.39	73.24	73.16	82.48
Min. F1 Score	63.10	63.41	57.63	54.55
Max. F1 Score	77.63	80.17	100.00	100.00
F1 Score SD	4.50	6.17	13.43	12.90

Tab. IV shows the average testing accuracy of the 'dis', 'low', 'mid', and 'hi' classifiers. We can see a monotonic increasing tendency in the accuracy as the classifier changes from 'low' to 'hi' from the average accuracy results. The average accuracy values of 'mid' and 'hi' classifiers were higher than 80%. Similar to the previous section of classifiers trained with different sets of features, an standard deviation of more than 5% can be observed with the classifiers that handled

different levels of discomfort. For the 'mid' and 'hi' classifiers, a standard deviation of more than 10% can be spotted.

VI. Discussions

In this study, the overall comfort level detection accuracy achieved with SVM was 71.88%, as is shown in Tab. III. For all the participants in this study, SVM reached an acceptable accuracy in the detection of human comfort level, achieving accuracy higher than 60% for all participants and accuracy higher than 70% for six participants in the 'comfort' vs. 'discomfort' case. Other than the ordinary accuracy in the generic case of 'comfort' vs. 'discomfort', SVM achieved good performance with the 'mid' and 'hi' classifiers with average accuracy values of 80.72% and 82.80% respectively. Moreover, except for the 100% accuracy values obtained with very small testing datasets, excellent performances were observed with Participant 2 ('mid' classifier, accuracy = 89.53%, 86 testing samples), Participant 3 ('hi' classifier, accuracy = 90.20%, 43 testing samples), Participant 6 ('mid' classifier, accuracy = 89.63%, 164 testing samples), and Participant 9 ('mid' classifier, accuracy = 98.90%, 182 testing samples). It indicates that the SVM classifier in this study can obtain average performance in general cases and achieve excellent performance in some specific cases.

Compared to some other similar studies [20]–[22] involving human emotion detection with various methods, the proposed approach in this study has obtained competitive and reasonable performance. Considering the differences in the recognition objects, adopted physiological signals, and approach/experimental designs, the comparison between the results cannot provide a definitive conclusion of the best approach to conduct a study of this kind. However, considering that human comfort is much more complex with a mixture of various affective states compared to emotions [48], the demonstrated results can still validate the feasibility of the human comfort detection methodology in this paper.

From both Tab. III and Tab. IV, a variation of no less than 5% can be spotted in the accuracy values across different participants. We believed that the major reason for the variation was that the levels of physiological arousal stimulated by the video stimuli were different for the participants. Different participants could have different sensitivity levels to the stimuli presented with the simulator. Participants with higher sensitivity to the stimuli would have gone through a more immersive experience during the experiment. More significant physiological arousal related to the changes in comfort levels would have been recorded and used to train and test the classifier. A better accuracy could be expected from these participants. And vice versa for the participants with lower sensitivity to the stimuli. The different arousal levels of physiological signals stimulated by the stimuli might be the factor behind the variance in the accuracy.

Some limitations with the current study still exist and will be discussed below. The participants in this study had a homogeneous demographic background. The participants might be more familiar with the autonomous driving technology given the engineering background, and thus the perception of comfort and discomfort in AVs for the participants might have been quite different from people without an engineering background. Also, only one female participant was within our participant group. To better validate the universality of the comfort detection approach, participants with different demographic backgrounds should be included in the research.

A sample size of 10 will usually be enough to obtain representative results for such studies, and existing similar studies [17], [49] also used similar sample sizes. However, it is true that we could obtain more comprehensive results with a larger sample size as an extension of this work in the future.

In the video stimuli, the hands of the virtual driver were still visible to the participants. To address this issue, the participants have been explained regarding the hands and instructed to ignore them and focus on the perception of vehicle driving behaviors. They were also instructed to report if the hands do have affected their comfort ratings. Because the hands only took a tiny portion at the bottom and were always attached to and moving with the steering wheel, they could be easily and intuitively considered part of the steering wheel. In addition, compared to the major stimuli occupying the majority of the screen and simulator motions, the influence of the hands was almost ignorable. This is also why we did not receive any feedback regarding the hands affecting the comfort ratings. However, it was still an imperfection in the stimuli. In the future, the hands should be eliminated from the visualization in the updated stimuli.

In the end, as a simulator-based study, the results may not perfectly reflect the situations in real vehicle studies. Although studies have been conducted to prove that both high-end [5] and low-cost [50] simulators could effectively generate experience close to real-world driving scenarios, differences may still exist between the experiences of being in a real vehicle and on a simulator. To reduce the gap between the simulator and real vehicle in this study, efforts have been spent on verbally instructing the participants to imagine the scenarios as well as tuning the scaling factors of the motions generated by the platform. Compared to the simulator used in [8] that successfully supported the research on human's situational awareness and trust in a semi-automated vehicle, we believed that our simulator with similar configurations could adequately fulfill the requirement of fidelity in this study. Besides, it is true that only by carrying out the experiment on an actual AV can we totally eliminate the gap, but the data collection and processing methods used in the simulator-based studies shall still be applicable to the real-vehicle-based studies. This is how the simulator-based studies retain the value even in the future when actual AVs are ready for conducting similar studies.

Considering that human physiological signals may incorporate noises that may have potentially impacted the detection performance, as the future work, one potential approach might help improve detection accuracy. We can combine the physiological-based classifier with a context-based comfort model to improve the accuracy. The measurement of physiological signals using the wearable sensing system can be affected by various factors, e.g. body movements and environmental temperature, and these factors are inevitable

if the system is used on a vehicle passenger. A lowered detection accuracy can result from these inevitable factors with physiological signal capture. To reduce the negative influence from the uncertainty of physiological signals, we can refer to the contextual information from the road or other vehicles in the stimuli to make corrections to the detection results merely from the physiological signals.

The normalization process has been conducted on the entire dataset for each participant. This has helped handle the variations across individuals. For the same individual, the physiological responses collected across different days have been normalized on the same basis. Our collected data showed that the same individual's responses usually did not vary much even on different days, so we think the normalization process could account for this variation effect.

We created the 'low-discomfort' vs. 'mid- & hi-discomfort', 'mid-discomfort' vs. 'low- & hi-discomfort', and the 'hi-discomfort' vs. 'low- & mid-discomfort' datasets and trained classifiers based on these datasets as an exploration of the possibility to perform multi-class detection of human comfort. With the existing SVM classifiers, the 'dis', 'low', 'mid', and 'hi' classifiers, a hierarchical structure multi-class SVM can be constructed [51]. The 'dis' classifier can act as the top-level classifier and determine whether a sample is a 'comfort' sample or a 'discomfort' sample. The sample classified as a 'discomfort' sample by the 'dis' classifier can be further classified as a 'discomfort' sample of a certain level after inputted into the next level classifier, which can be a One-Against-All multi-class SVM consisting of the 'low', 'mid', and 'hi' classifier. Considering that the accuracy of the individual classifiers was still not excellent at the current stage, no further step was taken in this study to test the idea of the hierarchical multi-class SVM. In the future, multi-class comfort level detection or even continuous comfort level detection can be further explored after the overall detection accuracy has been improved with more sensing signals and more sophisticated detection algorithms.

Different participants had different ranked lists of features in our study. This suggests that the classifier trained for one specific person would only work the best when being applied to this person. In some other studies [41], [52], a classifier with this individual-sensitive attribute can be defined as a subject-dependent classifier. To address this limitation and develop a classifier that is subject-independent is another challenging task.

VII. Conclusion

We have carried out a simulator-based study of human comfort subjected to autonomous vehicle ride scenarios. The driving simulator provided video and 6-DOF motion stimuli of AV rides to the participants. A wearable sensing system was used in the experiment to collect physiological data, and the subjective comfort level of the participants was captured with a press button. We employed an SVM classifier to detect the comfort level of the participants based on the data acquired during the experiment. According to the results, the experiment stimuli in the study successfully induced different levels of comfort in the participants. The protocol proposed in this study was proved to be feasible for studies on human comfort in AVs. The comfort level detection achieved a good overall accuracy. This proved the feasibility of using SVM to detect human comfort levels with physiological signals. Nevertheless, the accuracy varied a lot across different participants. In the future, we aim at further improving the accuracy of detection and extend the resolution of detection to a multi-class or continuous value measurement of the comfort level.

References

[1] T. Litman, *Autonomous Vehicle Implementation Predictions*. Victoria, BC, Canada: Victoria Transport Policy Institute, 2017.

[2] P. Fernandes and U. Nunes, "Platooning with IVC-enabled autonomous vehicles: Strategies to mitigate communication delays, improve safety and traffic flow," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 1, pp. 91–106, Mar. 2012.

[3] C. Rödel, S. Stadler, A. Meschtscherjakov, and M. Tscheligi, "Towards autonomous cars: The effect of autonomy levels on acceptance and user experience," in *Proc. 6th Int. Conf. Automot. User Interfaces Interact. Veh. Appl. (AutomotiveUI)*. New York, NY, USA: Association for Computing Machinery, Sep. 2014, pp. 1–8, doi: 10.1145/2667317.2667330.

[4] J. D. Power. (Oct. 23, 2019). *Mobility Pipe Dreams? J.D. Power and SurveyMonkey Uncover Shaky Consumer Confidence About the Future*. [Online]. Available: https://www.jdpower.com/business/press-releases/2019-mobility-confidence-index-study-fueled-surveymonkey-audience/

[5] H. Bellem *et al.*, "Can we study autonomous driving comfort in moving-base driving simulators? A validation study," *Hum. Factors*, vol. 59, no. 3, pp. 442–456, 2017.

[6] N. Du *et al.*, "Look who's talking now: Implications of AV's explanations on driver's trust, AV preference, anxiety and mental workload," *Transp. Res. C, Emerg. Technol.*, vol. 104, pp. 428–442, Jul. 2019.

[7] L. Petersen, H. Zhao, D. Tilbury, X. J. Yang, and L. P. Robert, "The influence of risk on driver's trust in semi-autonomous driving," in *Proc. Ground Vehicle Syst. Eng. Technol. Symp.*, Novi, MI, USA, 2018, pp. 1–10.

[8] L. Petersen, L. Robert, J. Yang, and D. Tilbury, "Situational awareness, driver's trust in automated driving systems and secondary task performance," *SAE Int. J. Connected Auton. Vehicles, Forthcoming*, vol. 2, p. 26, Mar. 2019, doi: 10.2139/ssrn.3345543.

[9] D. J. Oborne, "Vibration and passenger comfort," *Appl. Ergonom.*, vol. 8, no. 2, pp. 97–101, 1977.

[10] Y. Morales *et al.*, "Visibility analysis for autonomous vehicle comfortable navigation," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2014, pp. 2197–2202.

[11] M. P. De Looze, L. F. M. Kuijt-Evers, and J. Van Dieën, "Sitting comfort and discomfort and the relationships with objective measures," *Ergonomics*, vol. 46, no. 10, pp. 985–997, Aug. 2003.

[12] R. R. Bishu, M. S. Hallbeck, M. W. Riley, and T. L. Stentz, "Seating comfort and its relationship to spinal profile: A pilot study," *Int. J. Ind. Ergonom.*, vol. 8, no. 1, pp. 89–101, Aug. 1991.

[13] K. C. Parsons, E. M. Whitham, and M. J. Griffin, "Six axis vehicle vibration and its effects on comfort," *Ergonomics*, vol. 22, no. 2, pp. 211–225, 1979.

[14] M. J. M. Nor, M. H. Fouladi, H. Nahvi, and A. K. Ariffin, "Index for vehicle acoustical comfort inside a passenger car," *Appl. Acoust.*, vol. 69, no. 4, pp. 343–353, Apr. 2008.

[15] J. M. Devonshire and J. R. Sayer, "Radiant heat and thermal comfort in vehicles," *Hum. Factors, J. Hum. Factors Ergonom. Soc.*, vol. 47, no. 4, pp. 827–839, Dec. 2005.

[16] K. Ormuž and O. Muftić, "Main ambient factors influencing passenger vehicle comfort," in *Proc. 2nd Int. Ergonom. Conf. (Ergonomics)*, Stubičke Toplice, Croatia, 2004, pp. 1–8.

[17] A. P. De Vos, J. Theeuwes, W. Hoekstra, and M. J. Coëmet, "Behavioral aspects of automatic vehicle guidance: Relationship between headway and driver comfort," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 1573, no. 1, pp. 17–22, Jan. 1997.

[18] M. Elbanhawi, M. Simic, and R. Jazar, "In the passenger seat: Investigating ride comfort measures in autonomous cars," *IEEE Intell. Transp. Syst. Mag.*, vol. 7, no. 3, pp. 4–17, Jul. 2015.

[19] C. L. Lisetti and F. Nasoz, "Using noninvasive wearable computers to recognize human emotions from physiological signals," *EURASIP J. Adv. Signal Process.*, vol. 2004, no. 11, pp. 1672–1687, Dec. 2004.
[20] J. Kim and E. André, "Emotion recognition based on physiological changes in music listening," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 12, pp. 2067–2083, Dec. 2008.
[21] C. Maaoui and A. Pruski, "Emotion recognition through physiological signals for human-machine communication," *Cutting Edge Robot.*, vol. 2010, pp. 317–332, Sep. 2010.
[22] E.-H. Jang, B.-J. Park, M.-S. Park, S.-H. Kim, and J.-H. Sohn, "Analysis of physiological signals for recognition of boredom, pain, and surprise emotions," *J. Physiol. Anthropol.*, vol. 34, no. 1, pp. 1–12, Dec. 2015.
[23] Y. Yao, Z. Lian, W. Liu, C. Jiang, Y. Liu, and H. Lu, "Heart rate variation and electroencephalograph—The potential physiological factors for thermal comfort study," *Indoor Air*, vol. 19, no. 2, pp. 93–101, Apr. 2009.
[24] F. Salamone *et al.*, "Integrated method for personal thermal comfort assessment and optimization through users' feedback, IoT and machine learning: A case study," *Sensors*, vol. 18, no. 5, p. 1602, May 2018.
[25] M. Berger and L. Dörrzapf, "Sensing comfort in bicycling in addition to travel data," *Transp. Res. Proc*, vol. 32, pp. 524–534, Jan. 2018.
[26] M. Beggiato, F. Hartwich, and J. Krems, "Physiological correlates of discomfort in automated driving," *Transp. Res. F, Traffic Psychol. Behav.*, vol. 66, pp. 445–458, Oct. 2019.
[27] M. Beggiato, N. Rauh, and J. Krems, "Facial expressions as indicator for discomfort in automated driving," in *Proc. Int. Conf. Intell. Hum. Syst. Integr.* Cham, Switzerland: Springer, 2020, pp. 932–937.
[28] N. Dillen, M. Ilievski, E. Law, L. E. Nacke, K. Czarnecki, and O. Schneider, "Keep calm and ride along: Passenger comfort and anxiety as physiological responses to autonomous driving styles," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, Apr. 2020, pp. 1–13.
[29] M. G. Helander and L. Zhang, "Field studies of comfort and discomfort in sitting," *Ergonomics*, vol. 40, no. 9, pp. 895–915, Sep. 1997.
[30] F. Hartwich, M. Beggiato, and J. F. Krems, "Driving comfort, enjoyment and acceptance of automated driving—Effects of drivers' age and driving style familiarity," *Ergonomics*, vol. 61, no. 8, pp. 1017–1032, Aug. 2018.
[31] P. J. Gianaros, E. R. Muth, J. T. Mordkoff, M. E. Levine, and R. M. Stern, "A questionnaire for the assessment of the multiple dimensions of motion sickness," *Aviation, Space, Environ. Med.*, vol. 72, no. 2, pp. 115–119, 2001.
[32] T. G. Cengiz and F. C. Babalık, "An on-the-road experiment into the thermal comfort of car seats," *Appl. Ergonom.*, vol. 38, no. 3, pp. 337–347, May 2007.
[33] L. Zhang, M. G. Helander, and C. G. Drury, "Identifying factors of comfort and discomfort in sitting," *Hum. Factors, J. Hum. Factors Ergonom. Soc.*, vol. 38, no. 3, pp. 377–389, Sep. 1996.
[34] K. Lee Koay, M. L. Walters, and K. Dautenhahn, "Methodological issues using a comfort level device in human-robot interactions," in *Proc. IEEE Int. Workshop Robot Hum. Interact. Commun. (ROMAN)*, Nashville, TN, USA, Aug. 2005, pp. 359–364.
[35] Y.-P. Lin *et al.*, "EEG-based emotion recognition in music listening," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 7, pp. 1798–1806, Jul. 2010.
[36] B. Goldberg, K. W. Brawner, and H. K. Holden, "Efficacy of measuring engagement during computer-based training with low-cost electroencephalogram (EEG) sensor outputs," in *Proc. Hum. Factors Ergonom. Soc. Annu. Meeting*, vol. 56, no. 1. Los Angeles, CA, USA: Sage, 2012, pp. 198–202.
[37] P. Aspinall, P. Mavros, R. Coyne, and J. Roe, "The urban brain: Analysing outdoor physical activity with mobile EEG," *Brit. J. Sports Med.*, vol. 49, no. 4, pp. 272–276, 2015.
[38] C. Setz, B. Arnrich, J. Schumm, R. L. Marca, G. Tröster, and U. Ehlert, "Discriminating stress from cognitive load using a wearable EDA device," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 2, pp. 410–417, Mar. 2010.
[39] K. Kasos *et al.*, "Bilateral comparison of traditional and alternate electrodermal measurement sites," *Psychophysiology*, vol. 57, no. 11, Nov. 2020, Art. no. e13645, doi: 10.1111/psyp.13645.
[40] M. Benedek and C. Kaernbach, "A continuous measure of phasic electrodermal activity," *J. Neurosci. Methods*, vol. 190, no. 1, pp. 80–91, 2010.
[41] M. Ali, F. A. Machot, A. H. Mosa, and K. Kyamakya, "CNN based subject-independent driver emotion recognition system involving physiological signals for ADAS," in *Advanced Microsystems for Automotive Applications 2016*, T. Schulze, B. Müller, and G. Meyer, Eds. Cham, Switzerland: Springer, 2016, pp. 129–158.
[42] B. Figner and R. Murphy, "Using skin conductance in judgment and decision making research," in *A Handbook of Process Tracing Methods for Decision Research*, M. Schulte-Mecklenbeck, A. Kuehberger, and R. Ranyard, Eds. New York, NY, USA: Psychology Press, 2011, pp. 163–184.
[43] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 10, pp. 1175–1191, Oct. 2001.
[44] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, nos. 1–3, pp. 389–422, 2002.
[45] K. Duan, S. S. Keerthi, and A. N. Poo, "Evaluation of simple performance measures for tuning SVM hyperparameters," *Neurocomputing*, vol. 51, pp. 41–59, Apr. 2003.
[46] J. Snoek, H. Larochelle, and R. P. Adams, "Practical Bayesian optimization of machine learning algorithms," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 2951–2959.
[47] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Trans. Neural Netw.*, vol. 13, no. 2, pp. 415–425, Mar. 2002.
[48] D. Miller, *The Comfort of Things*. Cambridge, MA, USA: Polity, 2008.
[49] T. Yamakoshi *et al.*, "A preliminary study on driver's stress index using a new method based on differential skin temperature measurement," in *Proc. 29th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2007, pp. 722–725.
[50] D. Llopis-Castelló, F. J. Camacho-Torregrosa, J. Marín-Morales, A. M. Pérez-Zuriaga, A. García, and J. F. Dols, "Validation of a low-cost driving simulator based on continuous speed profiles," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2602, no. 1, pp. 104–114, Jan. 2016.
[51] L. Vanitha and G. R. Suresh, "Hierarchical SVM to detect mental stress in human beings using heart rate variability," in *Proc. 2nd Int. Conf. Devices, Circuits Syst. (ICDCS)*, Mar. 2014, pp. 1–5.
[52] J. Kim, "Bimodal emotion recognition using speech and physiological changes," in *Robust Speech Recognition and Understanding*, M. Grimm and K. Kroschel, Eds. Rijeka, Croatia: IntechOpen, 2007, ch. 15, doi: 10.5772/4754.

Haotian Su received the bachelor's degree from Tsinghua University, Beijing, China, in 2018. He is currently pursuing the Ph.D. degree in automotive engineering with Clemson University International Center for Automotive Research. He is currently involved in research related to autonomous driving and human factor studies.

Yunyi Jia (Senior Member, IEEE) received the B.S. degree from the National University of Defense Technology, Changsha, China, the M.S. degree from the South China University of Technology, Guangzhou, China, and the Ph.D. degree from Michigan State University, East Lansing, MI, USA. He is currently a McQueen Quattlebaum Assistant Professor and the Director of the Collaborative Robotics and Automation (CRA) Laboratory, Department of Automotive Engineering, Clemson University International Center for Automotive Research (CU-ICAR), Greenville, SC, USA. His research interests include robotics, autonomous vehicles, and advanced sensing systems.