






User Engagement with Driving Simulators: An Analysis of Physiological Signals

Ying-Hsang Liu ^{1,3}✉, Moritz Spiller ², Jinshuai Ma³, Tom Gedeon ³,
Md Zakir Hossain ³, Atiqul Islam³, and Ralf Bierig ⁴

¹ Department of Design and Communication, University of Southern Denmark,
Kolding, Denmark
`yingliu@sdu.dk`

² Medical Faculty/ University Clinic A.ö.R. (FME/UKMD),
Otto-von-Guericke-University Magdeburg, Magdeburg, Germany
`moritz.spiller@ovgu.de`

³ Research School of Computer Science, The Australian National University,
Canberra, Australia
<`ying-hsang.liu, jinshuai.ma, tom.gedeon, zakir.hossain,`
`atiqul.islam`>`@anu.edu.au`

⁴ Department of Computer Science, Maynooth University, Maynooth, Ireland
`ralf.bierig@mu.ie`

Abstract. Research on driving simulation has increasingly been concerned with the user's experience of immersion and realism in mixed reality environments. One of the key issues is to determine whether people perceive and respond differently in these environments. Physiological signals provide objective indicators of people's cognitive load, mental stress, and emotional state. Such data can be used to develop effective computational models and improve future systems. This study was designed to investigate the relationship between the verisimilitude of simple driving simulators and people's physiological signals, specifically GSR (galvanic skin response), BVP (blood volume pulse) and PR (pupillary response). A within-subject design user experiment with 24 participants for five different driving simulation environments was conducted. Our results reveal that there is a significant difference in the mean of GSR among the conditions of different configurations of simple driving simulators, but this is not the case for BVP and PR. The individual differences of gender, whether people wear glasses and previous experiences of driving a car or using a driving simulator are correlated with some physiological signals. The data is classified using a hybrid GA-SVM (genetic algorithm-support vector machine) and GA-ANN (artificial neural network) approach. The evaluation of the classification performance using 10-fold cross-validation shows that the choice of the feature subset has minor impact on the classification performance, while the choice of the classifier can improve the accuracy for some classification tasks. The results further indicate that the SVM is more sensitive to the selection of training and test data than the ANN. Our findings inform about the verisimilitude of simple driving simulators on the driver's perceived fidelity and physiological responses.

Implications for the design of driving simulators in support of training are discussed.

Keywords: Driving simulation · Virtual reality · Sensor · Eye tracking · User study

1 Introduction

Research on driving simulation is becoming increasingly interested in the user's experience of immersion and realism when these simulations are moved to virtual reality (VR) environments [32,37]. One of the key issues regarding realism in intelligent VR system design is to determine whether people perceive and respond differently in VR. In addition to subjective user-perceived measures that are extensively used in VR studies, physiological signals provide objective indicators of people's cognitive load, mental stress, and emotional state. Such data can be used to develop effective computational models and improve future systems. Some studies have used the physiological signals in VR environments [16,27], but the relationship between the features of user interfaces in VR environments and the physiological responses remains unclear.

Research on driving and flight simulation has been concerned with the issue of realism in virtual reality (VR) environments. Since the objective is to seek maximum realism, user perception issues, such as people's sense of presence (i.e., the feelings of being there) has been extensively studied [32,37]. Specifically, researchers have attempted to develop a driving simulator with an intelligent tutoring system, enhanced by a motion platform to improve presence [33]. However, in the context of flight simulation, the operator's perceived fidelity is not necessarily induced by the exact simulation of physical environments [32], and graphical fidelity alone is not correlated with galvanic skin responses (GSR) in the context of gaming [27]. One of the key issues regarding realism in intelligent VR system design is to determine whether people have different responses to the VR environments given user perceptions and physiological signals.

The use of physiological signals for building up computational models that can detect cognitive load, mental stress and emotional state for VR environments has the potentials for further development of user-adaptive interfaces. From user-centered design perspectives, the issues of user experience and physiological responses in VR environments have been emerging [11,29]. However, the issue of individual differences in cognitive processing and perception, which is important for developing user-adaptive interfaces, has received scant attention in driving and flight simulation studies [12,36]. More research on the effect of individual differences in cognitive processing and user perception will provide insights into the user-adaptive interface design in VR environments.

In our previous work we tried to establish a relationship between driving intervention task with simulator driving performance [14]. In this study we intend to address the issue of simulation validity by investigating the relationship between the verisimilitude of simple driving simulators and the physiological signals

of GSR, blood volume pulse (BVP) and pupillary response (PR). We construct computational models to detect physiological responses from an observer perspective and determine the relationship between the individual differences, user perceptions and the physiological responses in driving simulation. The specific research questions are as follows:

- Is there any difference in physiological responses for driving simulation environments?
- What is the relationship between the individual differences, user perceptions and the physiological responses in a driving simulation?
- To what extent computational models can detect different driving situations with high levels of accuracy?

Our key findings reveal that participants have significantly different GSR responses using a combination of the monitor, keyboard, driving set and VR headset of driving simulation environments. Individual differences such as gender, previous experiences of driving and using a simulator and user perceptions are correlated with some physiological responses. Our classification of the physiological data using a hybrid GA-SVM and GA-ANN approach can achieve a high level of accuracy, close to 90% for the driving situations of normal and emergency.

2 Related Work

2.1 User Issues in VR Environments

Research on driving and flight simulation in virtual reality environments has been concerned with user perception issues, such as people’s sense of presence and simulation sickness. To evaluate user experience in VR, the concept of presence (i.e., the feelings of being there) has been proposed and extensively used in the research literature [22,32]. For instance, the operator’s perceived fidelity (i.e., “the degree to which visual features in the virtual environment (VE) conform to visual features in the real environment” [32, p. 115]) is not necessarily induced by the exact simulation of physical environments and a sense of presence can be included in the formal assessment of fidelity. These user perception issues are important considerations for the design of user interfaces in VR environments.

In driving simulation environments vehicle velocity has been identified as a significant factor affecting driving simulation sickness and discomfort [25,34]. To provide driving skills training with the goal of improving presence and immersion in VR environments, a driving simulator with an intelligent tutoring system has been developed and evaluated [33]. These studies suggest that mental workload and user perception issues need further considerations for developing interfaces in VR environments.

2.2 User Perceptions and Physiological Responses

In addition to the use of questionnaires for assessing user perceptions, research has also been concerned with the user's cognitive and emotional states using physiological signals, such as skin conductance, heart rate and pupillary response. More specifically, physiological signals of skin conductance and heart rate have been suggested to assess emotional states as objective measures for people's responses in virtual environments [15]. Using a GSR sensor to measure physiological arousal, it was found that graphical fidelity is not correlated with GSR response [27]. To consider the effect of changes in scene brightness on the pupillary response for 2D screens and VR HMDs (head-mounted displays), an individual calibration procedure and constriction-based models of pupil diameter have been proposed [16]. However, the relationship between the features of user interfaces in VR environments and the responses measured by user-perceived data or physiological signals remains unclear.

From user-centered design perspectives, issues of usability of visualization systems, user experience and physiological responses in VR environments have received more attention in the research literature. For example, the user-centred design principles have been further applied to immersive 3D environments [11]. Besides, researchers have attempted to make connections between presence ratings and usability score [29], and between being present and levels of stress, measured by skin conductance response [37]. Overall, research has adopted the user-centered design principles and techniques for system design in VR environments.

2.3 Individual Differences

Aside from user perception issues, research on user experience in VR environments have touched on the issue of individual differences in cognitive processing and perception. For instance, it was found that there are gender differences in simulator sickness [25]. Females experience more simulator sickness than males. In the setting of a driving simulator, it was found that age makes a difference in user ratings for assistive technology [36], but there is no difference in attentional performance [6].

3 User Experiment

This study was designed to investigate the relationship between the verisimilitude of simple driving simulators and people's physiological signals, specifically galvanic skin response (GSR), blood volume pulse (BVP) and pupillary response (PR). A within-subject design user experiment with twenty-four participants was conducted for five configurations of driving simulation environment (See Figure 1) that used a combination of monitor(s), keyboard, driving set, and VR headset. The order of presentation was randomized by a Latin-squared design to minimize possible effects of learning and fatigue [19].

3.1 Apparatus

We used five configurations for our driving simulation as shown in Figure 1 and applied the driving simulator software, CCD ⁵ for driving environments.

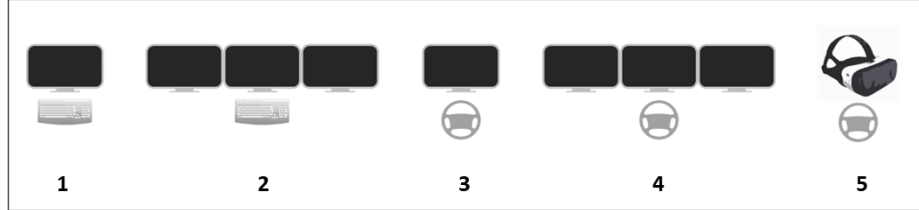


Fig. 1: The driving simulation environments include 1) single monitor and keyboard; 2) triple monitors and keyboard; 3) single monitor and driving set; 4) triple monitors and driving set; and 5) VR headset and driving set.

Since we were interested in people’s reactions to different driving environments, we chose to provide simple setups, steering, accelerating, braking and switching gears between forward and backward in automatic gear style. To collect data from both normal driving and emergency situations, the traffic and emergency levels in the CCD software were set to 70% that increased the likelihood of emergency situations like ‘Hit a car’. This was done based on our pilot study results

Table 1: Driving event with corresponding situations.

	Event	Situation
1	Hit a pedestrian	Emergency
2	Almost hit a pedestrian	
3	Hit an object	
4	Almost hit an object	
5	Hit a car	
6	Almost hit a car	
7	Normal driving	Normal
8	Stopping	

for stimulating experiences without getting bored or annoyed. While normal driving consisted of basic driving activity like moving forward and stopping, emergency situations included accidents or near-accidents involving objects, other vehicles and pedestrians as listed in Table 1. We implemented a manual labeling program for identifying these driving events that we describe in Section 4).

We used the E4 wristband ⁶ to collect real-time signals of GSR and BVP. A customized client program based on the E4 wristband API was developed for recording data, with a millisecond timestamp accuracy. We used the Eye-Tribe eye tracker ⁷ for acquiring pupil diameters with timestamps. To create

⁵ <https://citycardriving.com>

⁶ <https://www.empatica.com/en-gb/research/e4/>

⁷ <http://theeyetribe.com>

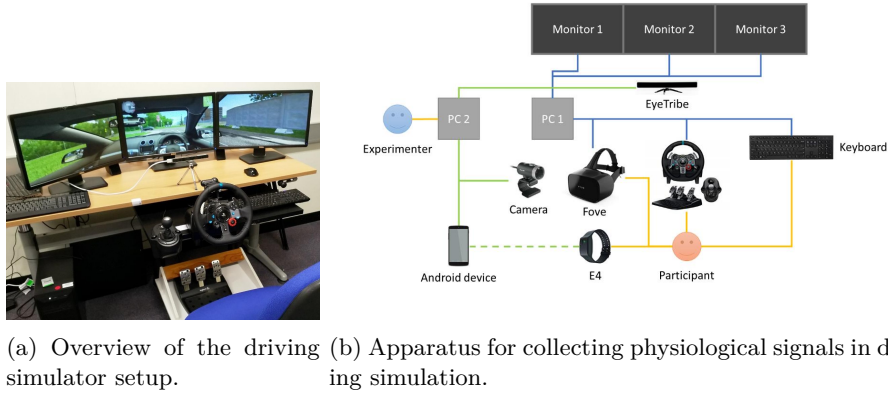


Fig. 2: Simulator setup and experimental apparatus

realistic VR environments, we used Fove VR headset⁸ with an integrated eye tracker. The main CCD was displayed inside the Fove VR headset, with a duplicated CCD window displayed in the central monitor for mouse operation. A customized client program based on Fove API was developed to record the pupillary response data. The user interface of the sensor program was displayed on the right monitor. All the physiological signals data were synchronised for data analysis (See Figure 2b).

3.2 Procedure

After a brief introduction and after consenting to the study⁹ the participant was instructed to wear the sensors that were then initialized and calibrated. A three-minute practice session allows the participant to become familiar with the devices, including the keyboard and steering wheel driving set. Then the participant was instructed to do free virtual driving for six minutes in each configuration using different devices, with two minutes breaks in-between. The experimenter ensured the proper setup at the beginning of each condition. Participants finished a total of five configurations of driving simulation environments (See Figure 1 and Figure 2a), followed by a questionnaire regarding demographic information, previous driving experiences and perceptions about the simulator.

4 Data Analysis

Our data analysis involved the labeling of each driving event in driving simulation environments, followed by the techniques of signal processing, feature

⁸ <https://www.getfove.com>

⁹ The study has been approved by the University Human Research Ethics Committee

extraction, feature selection and classification of physiological signals to predict the driving event, situations and experimental condition. Statistical analysis techniques were applied to examine the determine if there is any statistically significant difference by the driving situation, event, and experimental condition, as well as the relationship between user characteristics and physiological responses.

4.1 Labelling

During the experiment eight different driving events from two different categories can occur, as presented in Table 1. The CCD software logged each event that occurred during an experiment along with the related timestamp. By matching these log files with the files containing the recorded signals a labeled dataset as illustrated in Figure 3 was generated. Conclusively, the dataset is labeled by configuration, driving event and driving situation.

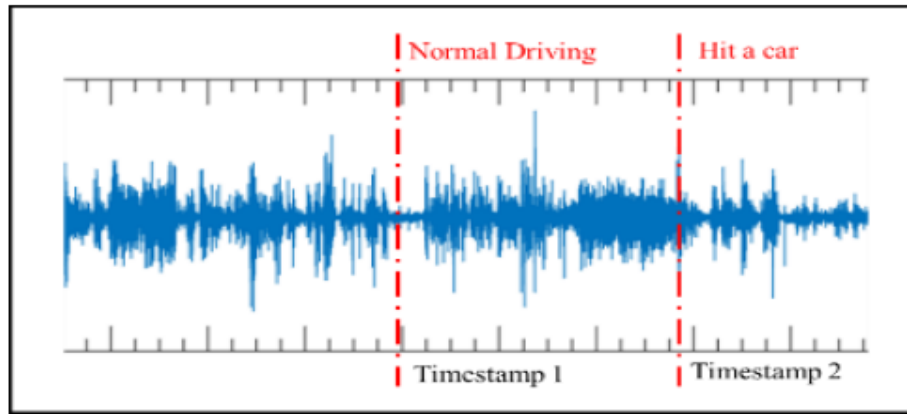


Fig. 3: Example of a labelled signal. The dashed lines mark an event and where obtained by matching the log files for the driving simulator software to the physiological signals using their timestamps.

4.2 Signal Preprocessing

The synchronized data was filtered to remove noise, which is consistently present in physiological signals recorded during user studies. This is caused by the external environment or movements of the participant. The Butterworth band-pass filter has been applied to both GSR and BVP signals [26]. The used bandpass for GSR signals was 0.1 Hz to 0.5 Hz [4], while the bandpass for BVP signals was 0.5 to 8 Hz [3,24]. The signal of the pupillary responses contains noise in the form of eye blinks, which cause a recorded pupil diameter of 0. To remove

those values linear interpolation was applied to the data [23], followed by the application of an S-G filter to smooth out the signal [35].

The individual participants may have different baselines in their physiological signals, which have to be removed by normalizing the measured data [5]. Max-Min Normalisation has been used to do that.

4.3 Feature Extraction

Segmentation In order to extract meaningful features from the recorded data, these data need to be segmented into subsegments of length $|n|$. This is done utilizing the event by which the data has been labeled and which have been introduced in Table 1. Figure 3 shows an example of a labeled GSR signal, where the red dotted lines correspond to that point in time when an event was logged. When an event was logged at time t_0 the interval $[t_0-1, t_0+2]$ has been extracted. This three-second data represent stimuli from which the features are calculated as described in Section 4.3. We chose this segmentation method according to the results of many preliminary studies where we observed the participant’s reaction to an event. On average, after two seconds, the participant’s physiological signal recovered from the stimuli. Due to a time delay between the actual stimuli and the point in time when the event is logged, we extract the data one second before the timestamp listed in the log.

Figure 4 illustrates a segmented GSR signal. While each red dotted line represents an event, the solid lines comprise the extracted time segment.

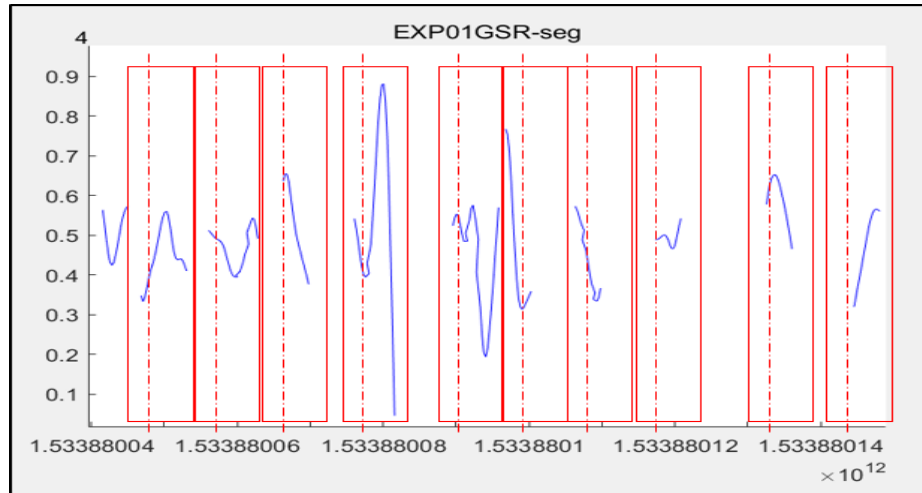


Fig. 4: Example of a segmented GSR signal. The solid lines comprehend one time segment of 3 seconds length, the dashed lines mark an event.

Feature Calculation Features were calculated from the time and the frequency domain [1,30]. To transform the time series data into the frequency domain, Fast Fourier Transform [10] was applied. We calculated mean absolute value (MAV), arithmetic mean (AM), root mean square (RMS), standard deviation (SD), waveform length (WL), zero crossing (ZC), skewness and kurtosis. Additionally, the absolute values of the recorded were summed up and the first and second difference between adjacent measurements were calculated for all signals. Zero crossing was calculated only for BVP signal since GSR and pupillary responses are not zero-mean. Waveform length was calculated for GSR and BVP signal. The skewness and the kurtosis are calculated on the frequency domain signal.

Since the classifier described in Section 4.5 are not scale invariant, the feature set was standardised to $[-1, 1]$.

4.4 Statistical Analysis

We construct mixed-effects models for determining the effects of driving simulators (condition) and events on physiological responses. Mixed-effects distinguish between fixed effects due to experimental condition and random effects due to individual differences in a sample. We choose the mixed-effects models because they are useful for the analysis of individual differences, with subjects and driving simulators as crossed random effects [2]. We use a logarithmic cross-ratio analysis [9] to determine if there is any significant relationship between individual differences and physiological responses.

4.5 Feature Selection & Classification

Genetic Algorithm The Genetic Algorithm (GA) is a commonly used feature selection method in machine learning applications [28] to optimize the performance of a classifier. GA is based on the "survival of the fittest" from Darwinian's evolution theory. It iteratively selects random feature subsets organized in populations and evaluates them on some fitness function. We used the classification accuracy of the respective classifier as the fitness function for the GA.

The size of a population has been set to ten, while the maximum number of generations was 1000. The four fittest feature combinations formed the next population by performing six mutations among them.

The termination criteria of the GA was the overall change in accuracy over the last ten iterations as presented in equation 1.

$$AccChange = x_n - \frac{\sum_{i=n-10}^n x_i}{n} \quad (1)$$

where $x \in X$ and $n = |X|$. X denotes the set of all calculated accuracies obtained from the respective classifier.

Support Vector Machine Support Vector Machines (SVMs) are a broadly used supervised classification algorithm [13,38]. A SVM classifies the data points by finding the best separating hyperplane in the n-dimensional feature space, which separates the data with the greatest margin possible.

For the experiments described in this paper an SVM model with a sigmoid kernel and an error rate of 5.0 has been used. These hyperparameters have been set after tuning the model using Grid Search.

Artificial Neural Network Artificial Neural Networks (ANNs) are supervised learning algorithms that are inspired by the working principle of the human brain and have been used successfully on physiological data [13]. ANNs consist of artificial neurons that are connected and are organized in layers. Each neuron can process received information and transmit it to the neurons connected to it. The signal is processed through the input layer, possibly followed by multiple hidden layers, to the output layer, which computes the final classification result.

The ANN used in our study consisted of one hidden layer with eight neurons, which used a Scaled Exponential Linear Unit (SELU) as activation function [20]. The output layer utilized the Softmax activation function and the weights were optimized using the Adam optimiser [18].

5 Results

This section reports the results and findings from the analysis of physiological responses in different situations, event and conditions, followed by the relationship between individual differences and the physiological responses. We then report the accuracy of classifying the physiological data, using a hybrid GA-SVM and GA-ANN approach for driving simulation environments.

5.1 Relationships among Event, Condition, and Physiological Signals

Our strategy for model fitting follows the approach by Baayen et al. [2]. Our null model initially includes random intercepts for condition and subject. To fit the data, we perform an automatic backward model selection of fixed and random parts of the linear mixed model [21]. Since the random intercepts for the subject are significant for both GSR and BVP, we choose a mixed-effects model with subject controlled as random effects.

Table 2 presents the constructed fixed and random effects models for both GSR and BVP. Model 1 is the baseline model with subject as random effects, whereas Models 2, 3, and 4 specify the fixed effects of the condition, event, a combination of both, as well as random effects.

Table 3 shows that Model 4 with condition and event as fixed effects accounts for 31.8% of variances, whereas Model 2 with the condition as fixed effects accounts for 21.4% of variances. Model 3 indicates that event as fixed effects only

Table 2: Model construction of fixed and random effects for measures of physiological responses by GSR and BVP.

Fixed and Random Effects Model	
Model 1	(1 subject)
Model 2	condition + (1 event) + (1 subject)
Model 3	event + (1 condition) + (1 subject)
Model 4	condition + event + (1 subject)

Note: Condition refers to types of driving simulator. Random intercepts for subject, event and condition are specified with (1|subject), (1|event) and (1|condition) respectively.

explain 0.6% of variances, though the effect of the event is statistically significant. Judging from the AIC value, Models 2, 3 and 4 are significantly better than our baseline Model 1. However, we cannot select the best model based on AIC alone since the values are close for Models 2, 3 and 4. Nonetheless, the results demonstrate that the event has significant but small effects on the mean of GSR. Condition, i.e. different configurations of the driving simulator has very significant effects on the mean of GSR.

Table 3: Model selection for effect of condition and event on mean of GSR.

	Mean of GSR			
	Model 1	Model 2	Model 3	Model 4
Condition		0.015*** (0.001)		0.015*** (0.001)
Event			-0.002* (0.001)	-0.002* (0.001)
Constant	0.041*** (0.003)	-0.003 (0.004)	0.051*** (0.012)	0.006 (0.006)
N	593	593	593	593
Log Likelihood	977.035	1,046.232	1,044.777	1,042.606
AIC	-1,948.069	-2,082.464	-2,079.554	-2,075.213
Marginal R^2	0.000	0.214	0.006	0.318
Conditional R^2	0.037	0.273	0.318	0.277

Note: *p < .05; **p < .01; ***p < .001; AIC: Akaike Information Criterion.

Further analysis reveals that condition, i.e. different configurations of driving simulator has significant effect on GSR, a measure of mental stress. Event has significant but small effect on GSR. Specifically, a configuration of the driving set and VR headset induces a higher level of stress than other configurations.

Table 4: Model selection for effect of condition and event on mean of BVP.

	Mean of BVP			
	Model 1	Model 2	Model 3	Model 4
Condition		0.003*** (0.001)		0.003*** (0.001)
Event			-0.001* (0.0004)	-0.001* (0.0004)
Constant	0.044*** (0.003)	0.035*** (0.003)	0.049*** (0.004)	0.040*** (0.004)
N	593	593	593	593
Log Likelihood	1,458.390	1,466.051	1,466.051	1,461.387
AIC	-2,910.781	-2,922.102	-2,922.102	-2,912.773
Marginal R^2	0.000	0.032	0.006	0.038
Conditional R^2	0.315	0.351	0.367	0.353

Note: * $p < .05$; ** $p < .01$; *** $p < .001$; AIC: Akaike Information Criterion.

Table 4 reveals that there is a very small effect of condition and event on the mean of BVP, a measure of cognitive load and emotional state. The fixed effects of condition only account for 3.2% of variances, whereas the effects of the event explain 3.8% of variances. The results suggest that the level of cognitive load measured by BVP does not change by condition and event. And we do not find statistically significant results for pupillary responses.

Overall, we find that participants have different levels of mental stress measured by GSR in different configurations of the driving simulator. The high level of stress is correlated with a configuration of the driving set and VR headset. The results suggest that the level of stress measured by GSR varies in different configurations. Therefore, our results confirm the validity of driving simulation in the simple setup with monitors and driving set. The use of VR headset has increased the level of stress, as observed in the significant differences in GSR signals.

5.2 Relationships among Individual Differences and Physiological Responses

To determine whether there is any correlation between the individual differences and the physiological signals, a logarithmic odds ratio analysis was conducted. Both dependent and independent variables were broken into “high” and “low” cases with the mean as cut point. Table 5 is a summary of the results. Overall, we find that the individual differences of gender, whether people wear glasses, user perceptions of devices affecting driving performance and whether people can see everything clearly through VR headset were correlated with the mean of GSR and BVP. That is, demographics, previous experience, and user perceptions are correlated with the GSR and BVP signals.

Table 5: Summary of the relationship between individual differences and physiological signals. User characteristics N = 593, Statistical significance at 95%.

	CPm	OR	LO	SE	t		CPm	OR	LO	SE	t
Gender						Driving simulator					
GSR mean	0.04	1.71	0.54	0.18	2.92*		0.04	0.48	-0.74	0.31	-2.39*
BVP mean	0.04	1.75	0.56	0.17	3.26*		0.04	1.02	0.02	0.25	0.08
Left eye PR	0.89	1.13	0.12	0.17	0.68		0.89	0.95	-0.05	0.26	-0.21
Right eye PR	4.32	1.87	0.63	0.17	3.65*		4.32	0.68	-0.39	0.26	-1.49
Wear glasses						Devices performance					
GSR mean	0.04	0.59	-0.53	0.20	-2.73*		0.04	0.69	-0.37	0.18	-2.07*
BVP mean	0.04	0.70	-0.36	0.18	-2.04*		0.04	0.51	-0.68	0.17	-3.93*
Left eye PR	0.89	1.07	0.07	0.18	0.40		0.89	0.92	-0.08	0.17	-0.47
Right eye PR	4.32	1.35	0.30	0.18	1.70		4.33	0.87	-0.14	0.17	-0.82
Driving license						VR Clearly					
GSR mean	0.04	0.45	-0.80	0.29	-2.72*		0.04	1.66	0.51	0.18	2.88*
BVP mean	0.04	0.79	-0.24	0.29	-0.83		0.04	1.63	0.49	0.17	2.96*
Left eye PR	0.89	1.17	0.16	0.30	0.53		0.89	1.22	0.20	0.17	1.18
Right eye PR	4.33	0.44	-0.82	0.30	-2.70*		4.33	1.23	0.21	0.16	1.26
Steering wheel											
GSR mean	0.04	1.15	0.14	0.21	0.66						
BVP mean	0.04	1.01	0.01	0.20	0.07						
Left eye PR	0.89	0.97	-0.03	0.20	-0.16						
Right eye PR	4.33	1.60	0.47	0.20	2.36*						

Note: CPm: Cut Points (Mean); OR: Odds Ratio; LO: Log Odds; SE: Standard Error; t: t-Value marked with asterisk (*) when statistically significant.

Specifically, female participants were more likely to have a higher mean of GSR and BVP than male participants. People who wear glasses were more likely to have a lower mean of GSR and BVP than people who didn't wear glasses. People who feel that devices affecting driving performance were more likely to have a lower mean of GSR and BVP. People who feel that they can see everything clearly through VR headset were more likely to have a higher mean of GSR and BVP. People who have more experiences using a driving simulator were more likely to have a lower mean of GSR, a measure of mental stress.

In other words, female participants were more likely to have higher levels of stress. Participants with more previous experiences (i.e., wearing glasses, driving license, driving simulator) were more likely to have lower levels of stress. The gender differences were also found in the BVP signals, a measure of cognitive load and emotional state. Female participants were more likely to have a higher mean of BVP by a factor of 1.75, or 75% than male participants.

Concerning the pupillary responses, female participants and those who have more left steering wheel experiences were more likely to have a higher mean of right eye pupillary response by a factor of 1.87 (or 87%) and 1.60 (or 60%) respectively. By contrast, people who have a driving license were more likely to

have a lower mean of GSR by a factor of 0.45 (or 55%), and have a lower mean of right eye pupillary response by a factor of 0.44 (or 56%).

These results suggest that gender is an important demographic factor affecting participants' physiological responses to different environments in driving simulation. It is more likely that female participants are more stressed and use more cognitive resources in the simulation environment. Participants with previous experiences in real-life driving and exposure to driving simulators are more likely to have a lower level of stress. Participants' perceptions about whether the devices affect their performance and whether they can see clearly through the VR headset are correlated with GSR and BVP in opposite directions.

5.3 Classification

The ANN and SVM described in the Sections 4.5 and 4.5 were trained on the 34 features described in Section 4.3. Additionally, we included the participants' gender in the feature set because our results suggest that gender is correlated with physiological responses in this study (See Table 5). Consequently, the final dataset included 35 features.

Tables 6a and 6b describe the accuracy (Acc) and standard deviation (SD) obtained by applying 10-fold cross-validation. Iterations correspond to the number of necessary iterations the GA required until the optimal performance was reached. The presented number of iterations corresponds to the total number of iterations until the GA terminated. The termination criteria were the overall change in accuracy over the last 10 iterations as presented in equation 1.

Tables 6a and 6b summarize the classification performance of the GA-SVM and GA-ANN approach. The high standard deviation using GA-SVM suggests, that the SVM is more sensitive to the distribution of the data samples in the training and test set.

The SVM shows significantly better performance when classifying the condition (5 classes, e.g. "VR headset") in which the participant is driving. However, since its standard deviation is much higher than the ANN's standard deviation, this result largely depends on the distribution of the data samples.

The ANN outperforms the SVM when classifying the event (8 classes, e.g. "hit a pedestrian" or "stopping"). As again indicated by the standard deviation of the SVM, it might be able to match the ANNs performance at this classification task.

Both classifiers perform similarly when classifying the driving situation (2 classes, "normal" or "emergency").

The GA required only a few iterations to find the optimal feature subset on both classification approaches. This indicates that the selection of a feature subset has only a minor impact on the classification performance of the used classifiers. However, both classifiers pursued to reduce the feature set, which initially included 35 features. The GA-ANN approach selected 22 features for the condition, 18 features for the event and 16 features for the situation classification task. The GA-SVM selected 20, 14 and 17 features respectively.

Table 6: Classification results.

	Acc [%]	SD	Iterations		Acc [%]	SD	Iterations
Condition	59.33	12.25	26	Condition	47.09	0.16	10
Event	65.87	6.46	10	Event	71.24	1.87	12
Situation	87.49	2.28	11	Situation	87.75	1.02	10

(a) GA-SVM

(b) GA-ANN method

6 Discussion

6.1 Is There any Difference in Physiological Responses for Driving Simulation Environments?

Our findings indicate that there is no significant difference in participants' BVP and pupillary responses in the configurations of driving environments. However, there are significant differences in GSR. Since the environmental simulations can be validated by the ability to replicate human responses in physical environments [22], our study suggests that people do not have different responses to the driving simulations by BVP signals and pupillary responses. Our results partially support the use of simple driving simulators as empirical tools in user behavior research. Our finding that there are significant differences in GSR for different driving simulators, however, shows that the use of VR headset induces a higher level of physiological arousal. In the context of virtual environments intended to create the feeling of presence in immersive environments, research shows that increased graphical quality alone is not correlated with GSR responses in the gaming context [15]. Therefore, our findings suggest that participants do not have different BVP and pupillary responses, while the use of a VR headset and a steering wheel driving set induces a higher level of physiological arousal in driving simulation environments.

6.2 What is the Relationship Between the Individual Differences, User Perceptions and the Physiological Responses in a Driving Simulation?

Our findings show that gender is an important factor affecting physiological responses to different environments in driving simulation. Previous research on the role of gender in simulator sickness has been inconclusive [25,36]. Our results support that females experience more simulator sickness than males [25], and females are more likely to feel stressed and use more cognitive resources in the simulation environment. Participants with previous experiences in real-life driving and exposure to driving simulators are found to have a lower level of stress. These findings correspond to the results on driving style familiarity and driving comfort [12], showing that driving style familiarity interacts with driving comfort by different age groups. Therefore, the demographic variables of gender,

previous driving experience and age and their effects on physiological responses and user perceptions need further research.

Concerning user perceptions, we find that there are discrepancies between the perceived feelings and physiological responses. Specifically, whether the devices affect participants' performance and whether they can see clearly through the VR headset are correlated with GSR and BVP responses in opposite directions. We speculate that since the use of a VR headset induces a higher level of physiological arousal and participants are engaged with the experiment, they are more likely to have higher GSR and BVP responses when prompted with the question of whether they can see clearly through VR headset. On the other hand, user perceptions about their performance might be explained by the new simulation environments introduced in our user experiment. Nonetheless, future research needs to consider user perceptions of speed and distance in simulated environments [17,22,33].

6.3 To What Extent can Computational Models Detect Different Driving situations with High Levels of Accuracy?

Our findings suggest that classifying the physiological data using a hybrid GA-SVM and GA-ANN approach can achieve a high level of accuracy, close to 90% for driving situations. Our performance evaluation using 10-fold cross-validation shows that the choice of the feature subset has minor impact on the classification performance, while the choice of the classifier can improve the accuracy for some classification tasks.

The results described in Section 5.3 and shown in the Tables 6 suggest the Genetic Algorithm mainly converged after few iterations. Therefore, it seems that the choice of a specific feature subset has a minor impact on the performance of the classifier. Nevertheless, the SVM classifier required more iterations and shows higher standard deviation, which corroborates the observations of other researchers [28] that the SVM is more sensitive to the features used for training.

Our research shows that it is possible to detect what kind of peripheral devices the user applies during the usage of a driving simulator or similar software. That contributes to the development of user adapted simulator software or games. By detecting the type of peripheral device that is used to visualize the software, the resolution or the layout of the user interface can be adapted to the specific device like a VR-Headset. In the case of a driving simulator, the configuration of the driving parameters can be adapted to a keyboard or a driving set. Among others, the latency and accuracy of how the steering impulses of the user are processed by the software can be altered.

6.4 Limitations and Future Research

Since this user experiment was conducted in a laboratory setting, one should be cautious about the generalisability of the results to the general population.

Our segmentation method described in Section 4.3 works based on the labels that have been matched with logged events from physiological signals. Due to

the nature of the domain and the driving simulator experiment in particular as normal driving situations will always occur much more frequently than emergencies. This leads to multiple subsequent labels for normal driving compared to other labels in the dataset, resulting in the trained model to be biased towards normal driving with good classification results for this class and reduced results for less frequent driving events. Researchers [8,31] have proposed two approaches for minority oversampling to improve learning from imbalanced datasets that we will apply in future work.

For the feature selection of eye gaze data, we hypothesise that the number of fixations, as well as the average fixation duration, are higher during an emergency event like "Hit a pedestrian" than in a normal driving situation. At the same time, saccades are expected to occur more frequently during normal driving situations. We suggest future research on additional features (e.g. fixations and saccades) and user-perceived sensory fidelity in different simulation environments for enhancing the user experience of presence [27,29] to make different driving situations more distinctive.

The analysis of pupillary responses could be enhanced by further considering scene brightness caused by the changes in driving scenes in VR environments [16]. To validate the use of the simulated environment for driving skills training purposes, future research can use new driving scenarios in automated driving simulators and simulation of tactile or audio feedback of the real driving environment, with particular emphasis on the usability issues through the analysis of physiological signals [7,29]. Additionally, the sensitivity of the SVM to physiological signal processing can be further investigated.

Our findings inform about the verisimilitude of simple driving simulators on the driver's perceived fidelity and physiological responses. This can be used to inform on the design of driving simulators in support of training. We suggest that individual differences such as prior driving experiences need to be considered in the design of a driving simulator (e.g. by offering difficulty-levels). Virtual environments can increase immersion while also increasing stress levels that should be considered in design (e.g. by leveraging realism through adjusting the likelihood of potentially dangerous situations).

7 Conclusion

We investigated the relationship between the verisimilitude of simple driving simulators and people's physiological signals, specifically galvanic skin response (GSR), blood volume pulse (BVP) and pupillary response (PR). We found that participants do not have different BVP and PR in driving simulation environments, which supports the use of steering wheel driving set as empirical tools in user behavior research. Individual differences such as previous experiences of driving should be considered in the design of driving simulators since they are correlated with physiological responses. In terms of predictability, our results further suggest that classifying the physiological data using a hybrid GA-SVM and GA-ANN approach can achieve a high level of accuracy, close to 90% for

driving situations while showing that the choice of the feature subset only has a minor impact on the classification performance. Our findings inform about the verisimilitude of simple driving simulators on the driver's perceived fidelity and physiological responses and provide implications for the design of future driving simulators.

References

1. Ayata, D., Yaslan, Y., Kamasak, M.: Emotion Recognition via Galvanic Skin Response: Comparison of Machine Learning Algorithms and Feature Extraction Methods. *Istanbul University - Journal of Electrical and Electronics Engineering* **17**(1), 3129–3136 (2017)
2. Baayen, R.H., Davidson, D.J., Bates, D.M.: Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* **59**(4), 390–412 (2008). <https://doi.org/10.1016/j.jml.2007.12.005>
3. Bagha, S., Hills, S., Bhubaneswar, P., Shaw, L.: A Real Time Analysis of PPG Signal for Measurement of SpO₂ and Pulse Rate. *International Journal of Computer Applications* **36**(11), 975–8887 (2011). <https://doi.org/10.5120/4537-6461>
4. Boucsein, W.: *Electrodermal Activity*. Springer, New York, 2nd edn. (2012). <https://doi.org/10.1007/978-1-4614-1126-0>
5. Cacioppo, J.T., Rourke, P.A., Marshall-Goodell, B.S., Tassinary, L.G., Baron, R.S.: Rudimentary physiological effects of mere observation. *Psychophysiology* **27**(2), 177–186 (1990). <https://doi.org/10.1111/j.1469-8986.1990.tb00368.x>
6. Cassarino, M., Maisto, M., Esposito, Y., Guerrero, D., Chan, J.S., Setti, A.: Testing Attention Restoration in a Virtual Reality Driving Simulator. *Frontiers in psychology* **10**, 250 (2019). <https://doi.org/10.3389/fpsyg.2019.00250>
7. Dols, J.F., Molina, J., Camacho, F.J., Marín-Morales, J., Pérez-Zuriaga, A.M., Garcia, A.: Design and Development of Driving Simulator Scenarios for Road Validation Studies. *Transportation Research Procedia* **18**, 289–296 (2016). <https://doi.org/10.1016/j.trpro.2016.12.038>
8. Douzas, G., Bacao, F.: Effective data generation for imbalanced learning using conditional generative adversarial networks. *Expert Systems with Applications* **91**, 464–471 (2018). <https://doi.org/10.1016/j.eswa.2017.09.030>
9. Fleiss, J.L., Levin, B., Paik, M.C.: *Assessing Significance in a Four-fold Table*. John Wiley & Sons, Hoboken, NJ, 3rd edn. (2003). <https://doi.org/10.1002/0471445428.ch3>
10. Frigo, M., Johnson, S.G.: FFTW: an adaptive software architecture for the FFT. In: *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No. 98CG36181)*. pp. 1381–1384 (1998)
11. Gerjets, P., Lachmair, M., Butz, M.V., Lohmann, J.: Knowledge Spaces in VR: Intuitive Interfacing with a Multiperspective Hypermedia Environment. In: *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. pp. 555–556 (2018). <https://doi.org/10.1109/VR.2018.8446137>
12. Hartwich, F., Beggiato, M., Krems, J.F.: Driving comfort, enjoyment and acceptance of automated driving—effects of drivers' age and driving style familiarity. *Ergonomics* **61**(8), 1017–1032 (2018). <https://doi.org/10.1080/00140139.2018.1441448>

13. Hossain, M.Z., Gedeon, T.: Observers' physiological measures in response to videos can be used to detect genuine smiles. *International Journal of Human-Computer Studies* **122**(November 2017), 232–241 (2019). <https://doi.org/10.1016/j.ijhcs.2018.10.003>
14. Islam, A., Ma, J., Gedeon, T., Hossain, M.Z., Liu, Y.: Measuring user responses to driving simulators: A galvanic skin response based study. In: 2019 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR). pp. 33–40 (2019). <https://doi.org/10.1109/AIVR46125.2019.00015>
15. Jang, D.P., Kim, I.Y., Nam, S.W., Wiederhold, B.K., Wiederhold, M.D., Kim, S.I.: Analysis of Physiological Response to Two Virtual Environments: Driving and Flying Simulation. *CyberPsychology & Behavior* **5**(1), 11–18 (2002). <https://doi.org/10.1089/109493102753685845>
16. John, B., Raiturkar, P., Banerjee, A., Jain, E.: An Evaluation of Pupillary Light Response Models for 2D Screens and VR HMDs. In: Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology. pp. 19:1–19:11. VRST '18, ACM, New York (2018). <https://doi.org/10.1145/3281505.3281538>
17. Kemeny, A., Panerai, F.: Evaluating perception in driving simulation experiments. *Trends in Cognitive Sciences* **7**(1), 31–37 (2003). [https://doi.org/10.1016/S1364-6613\(02\)00011-6](https://doi.org/10.1016/S1364-6613(02)00011-6)
18. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization. In: Proceedings of the 3rd International Conference for Learning Representations (2015)
19. Kirk, R.E.: Experimental design: Procedures for the behavioral sciences. Brooks/Cole, Pacific Grove, CA, 4th edn. (2013). <https://doi.org/10.4135/9781483384733>
20. Klambauer, G., Unterthiner, T., Mayr, A., Hochreiter, S.: Self-Normalizing Neural Networks. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) *Advances in Neural Information Processing Systems* 30. pp. 971–980. Curran Associates, Inc. (2017)
21. Kuznetsova, A., Brockhoff, P.B., Christensen, R.H.B.: lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software* **82**(13), 1–26 (2017). <https://doi.org/10.18637/jss.v082.i13>
22. Marín-Morales, J., Higuera-Trujillo, J.L., De-Juan-Ripoll, C., Llinares, C., Guixeres, J., Iñarra, S., Alcañiz, M.: Navigation Comparison between a Real and a Virtual Museum: Time-dependent Differences using a Head Mounted Display. *Interacting with Computers* (2019). <https://doi.org/10.1093/iwc/iwz018>
23. Mathôt, S., Dalmaijer, E., Grainger, J., Van der Stigchel, S.: The pupillary light response reflects exogenous attention and inhibition of return. *Journal of vision* **14**(14), 7 (2014). <https://doi.org/10.1167/14.14.7>
24. Mohd-Yasin, F., Yap, M.T., Reaz, M.B.I.: CMOS instrumentation amplifier with offset cancellation circuitry for biomedical application. *WSEAS Transactions on Circuits and Systems* **6**(1), 171–174 (2007)
25. Mourant, R.R., Thattacherry, T.R.: Simulator Sickness in a Virtual Environments Driving Simulator. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* **44**(5), 534–537 (2000). <https://doi.org/10.1177/154193120004400513>
26. Nabian, M., Yin, Y., Wormwood, J., Quigley, K.S., Barrett, L.F., Ostadabbas, S.: An open-source feature extraction tool for the analysis of peripheral physiological data. *IEEE Journal of Translational Engineering in Health and Medicine* **6**, 2800711 (2018). <https://doi.org/10.1109/JTEHM.2018.2878000>
27. Ocasio-De Jesús, V., Kennedy, A., Whittinghill, D.: Impact of Graphical Fidelity on Physiological Responses in Virtual Environments. In: Proceedings of the 19th

- ACM Symposium on Virtual Reality Software and Technology. pp. 73–76. VRST '13, ACM, New York (2013). <https://doi.org/10.1145/2503713.2503751>
28. Paiva, J.S., Cardoso, J., Pereira, T.: Supervised learning methods for pathological arterial pulse wave differentiation: A SVM and neural networks approach. *International Journal of Medical Informatics* **109**, 30–38 (2018). <https://doi.org/10.1016/j.ijmedinf.2017.10.011>
 29. Pettersson, I., Karlsson, M., Ghiurau, F.T.: Virtually the Same Experience?: Learning from User Experience Evaluation of In-vehicle Systems in VR and in the Field. In: *Proceedings of the 2019 on Designing Interactive Systems Conference*. pp. 463–473. DIS '19, ACM, New York (2019). <https://doi.org/10.1145/3322276.3322288>
 30. Phinyomark, A., Limsakul, C., Phukpattaranont, P.: A Novel Feature Extraction for Robust EMG Pattern Recognition. *Journal of Medical Engineering & Technology* **1**(1), 71–80 (2009). <https://doi.org/10.3109/03091902.2016.1153739>
 31. Piri, S., Delen, D., Liu, T.: A synthetic informative minority over-sampling (SIMO) algorithm leveraging support vector machine to enhance learning from imbalanced datasets. *Decision Support Systems* **106**, 15–29 (2018). <https://doi.org/10.1016/j.dss.2017.11.006>
 32. Robinson, A., Mania, K.: Technological research challenges of flight simulation and flight instructor assessments of perceived fidelity. *Simulation & Gaming* **38**(1), 112–135 (2007). <https://doi.org/10.1177/1046878106299035>
 33. Ropelato, S., Zund, F., Magnenat, S., Menozzi, M., Summer, R.W.: Adaptive Tutoring on a Virtual Reality Driving Simulator. *International SERIES on Information Systems and Management in Creative eMedia (CreMedia)* pp. 12–17 (2018)
 34. Sakamura, Y., Tomita, A., Shishido, H., Mizunami, T., Inoue, K., Kameda, Y., Harada, E.T., Kitahara, I.: A Virtual Boarding System of an Autonomous Vehicle for Investigating the Effect of an AR Display on Passenger Comfort. In: *2018 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. pp. 344–349 (2018). <https://doi.org/10.1109/ISMAR-Adjunct.2018.00101>
 35. Schafer, R.: What Is a Savitzky-Golay Filter? [Lecture Notes]. *IEEE Signal Processing Magazine* **28**(4), 111–117 (2011). <https://doi.org/10.1109/MSP.2011.941097>
 36. Schultheis, M.T., Rebimbas, J., Mourant, R., Millis, S.R.: Examining the Usability of a Virtual Reality Driving Simulator. *Assistive Technology* **19**(1), 1–10 (2007). <https://doi.org/10.1080/10400435.2007.10131860>
 37. Skarbez, R., Brooks Jr., F.P., Whitton, M.C.: Immersion and Coherence in a Stressful Virtual Environment. In: *Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology*. pp. 24:1–24:11. VRST '18, ACM, New York (2018). <https://doi.org/10.1145/3281505.3281530>
 38. Wu, Y., Liu, Y., Tsai, Y.H.R., Yau, S.T.: Investigating the role of eye movements and physiological signals in search satisfaction prediction using geometric analysis. *Journal of the Association for Information Science and Technology* (2019). <https://doi.org/10.1002/asi.24240>