Predicting Takeover Performance in Conditionally Automated Driving

Na Du
University of Michigan
Ann Arbor, MI 48109, USA
nadu@umich.edu

Dawn M. Tilbury
Lionel P. Robert
University of Michigan
Ann Arbor, MI 48109, USA
tilbury, lprobert@umich.edu

Feng Zhou
University of Michigan
Dearborn, MI 48128, USA
fezhou@umich.edu

Anuj K. Pradhan
University of Massachusetts
Amhert
Amherst, MA 01002, USA
anujkpradhan@umass.edu

Elizabeth Pulver
State Farm Mutual Automobile Insurance Company
Bloomingtom, IL 61710, USA
elizabeth.pulver.rr2@statefarm.com

X. Jessie Yang
University of Michigan
Ann Arbor, MI 48109, USA
xijyang@umich.edu

Abstract
In conditionally automated driving, drivers decoupled from operational control of the vehicle have difficulty taking over control when requested. To address this challenge, we conducted a human-in-the-loop experiment wherein the drivers needed to take over control from an automated vehicle. We collected drivers’ physiological data and data from the driving environment, and based on which developed random forest models for predicting drivers’ takeover performance in real time. Drivers’ subjective ratings of their takeover performance were treated as the ground truth. The best random forest model had an accuracy of 70.2% and an F1-score of 70.1%. We also discussed the implications on the design of an adaptive in-vehicle alert system.

Author Keywords
Transition of control; predictive modeling; adaptive in-vehicle alert system, human-automation interaction; human-robot interaction

CCS Concepts
• Human-centered computing → Empirical studies in HCI;

Introduction
According to the Society of Automotive Engineers [23], driving automation ranges from L0 (no automation) to L5 (full
automation). Automated driving features at SAE Level 3, such as the Audi A8’s Traffic Jam Pilot, are expected to be introduced into the market in 2020. With SAE Level 3 conditional automation, drivers serve as a fallback for the automation but are no longer required to actively monitor the driving environment and can engage in non-driving-related tasks (NDRTs). The concern from a safety perspective is that the limited driver-vehicle interaction puts drivers out-of-the-loop and increases the difficulty for drivers to take over control of the vehicle when the automated vehicle reaches its operational limits [2, 10, 11, 19, 27].

In order to resolve such difficulty, a substantial amount of research has been conducted on the factors that influence drivers’ takeover performance, including the driving environment [11, 15], types of NDRTs [9, 24, 26], and individual characteristics [7, 12, 26]. However, these studies emphasize on the high-level relationships between certain factors and takeover performance (e.g., elderly drivers have more difficulty in takeover transitions compared to younger drivers). It is still difficult to predict whether a particular driver is able to takeover successfully at a particular event. Therefore, computational models that are capable of predicting drivers’ takeover performance in real time are needed.

The advances in eye tracking and wearable sensors technology make it convenient to collect drivers’ physiological signals to reflect drivers’ states and interactions with environments. To our knowledge, only one study has used machine learning to predict takeover performance using drivers’ gaze behaviors and situation complexity [4]. In contrast to [4], our study aimed to predict drivers’ takeover performance when they are engaged in the same type of NDRT with different levels of cognitive load.

In this paper, we conducted a human-subject experiment wherein drivers experienced various takeover scenarios in a high-fidelity driving simulator. We collected drivers’ galvanic skin response (GSR), heart rate (HR), and eye movements were recorded for model development as they have been used as valid tools to assess drivers’ cognitive states and situational awareness of the driving environments [16, 21, 25]. Using drivers’ physiological data and environment parameters, we developed random forest models that were able to predict drivers’ takeover performance with the accuracy and F1-score both over 70%. The model can be used to guide the design of an adaptive in-vehicle alert system to improve takeover performance in conditionally automated driving.

Method
Participants
A total of 102 university students participated in the study (mean age = 22.9; SD = 3.8; range = 18-38; 40 females; 62 males). All the participants had normal or corrected-to-normal vision and a valid driver’s license.

Apparatus and stimuli
The study was conducted in a fixed-based driving simulator from Realtime Technologies Inc. (RTI, Michigan). The virtual world was projected on three front screens (16 feet away), one rear screen (12 feet away), and two side mirror displays (Figure 1).

This simulator was equipped with the SmartEye four-camera eye-tracking system (Smart Eye, Sweden) (Figure 2). The sampling rate of the eye-tracking system is 120 Hz. The Shimmer3 GSR+ unit (Shimmer, MA, USA) including GSR electrodes and PPG probes was used to collect GSR and HR data with a sampling rate of 128 Hz (Figure 3). The iMotions software (iMotions, MA, USA) was used for physi-
logical data synchronization and visualization in real-time (Figure 4).

The simulated vehicle was controlled by a steering wheel and pedal system embedded in a Nisan Versa car model. The vehicle was programmed to simulate an SAE Level 3 automation. Participants could press the button on the steering wheel to activate the automated mode. Once the AV reached its performance limit, an auditory TOR (“Takeover”) was issued, and the automated mode would be deactivated simultaneously for drivers to take over control of the vehicle.

The NDRT was a visual N-back memory task, adapted from the study of [13]. A series of stimuli were presented for 500 ms each in sequence with a 2500 ms interval. Participants were required to press the “Hit” button when the current stimulus was the same as the one presented N steps back in the sequence and press the “Reject” button otherwise. The task was running on an 11.6-inch touch screen tablet mounted in the vehicle (Figure 6).

**Experimental design**

The study employed a within-subjects design with the driver’s cognitive load, traffic density, and TOR lead time as independent variables. The cognitive load was manipulated via the difficulty of the NDRTs (low: 1-back memory task; high: 2-back memory task). The heavy and light traffic conditions had 15 and 0 oncoming vehicles per kilometer respectively [11]. The TOR lead time was 4 or 7 seconds [10]. Based on prior literature [14, 17, 18, 22], eight takeover events were designed in urban and rural drives with typical roadway features: 1) bicyclists ahead; 2) construction zone on the left; 3) construction zone ahead; 4) sensor error on the right curve; 5) swerving vehicle ahead; 6) no lane markings; 7) sensor error on the left curve; 8) police vehicle on shoulder. The order of cognitive load, traffic density and TOR lead time was counterbalanced via an $8 \times 8$ balanced Latin Square across participants. Considering standard programming practices for the simulator, the order of scenario presentations was counterbalanced by having half of the participants drive from Event 1 to 8, and the other half from Event 8 to 1.

**Experimental procedure**

After participants signed an informed consent form and completed an online demographics questionnaire, they were fitted with the eye-tracker and two GSR electrodes on left foot and the PPG probe on the left ear lobe. Participants had a 2-minute practice for the N-back memory task and a 5-minute practice drive to get familiar with the simulator environment. Next, each participant drove two experimental drives (15-20 minutes each), each containing four takeover events. At the beginning of the drive, participants were asked to activate the AV mode and then start the N-back task when the audio command “Please start the NDRT” was issued. Participants were informed that they would get additional 20 dollars if their NDRT performance was ranked top 10. After performing the NDRT for about 90-seconds, a TOR was issued unexpectedly, and participants were required to terminate the NDRT manually and take over control immediately. Participants were allowed to re-active the AV mode once they thought they had completed the takeover. After each takeover event, participants reported their takeover performance for that takeover event using a visual analogue scale on the tablet, with 0 indicating not good at all and 100 indicating very good (Figure 5).

**Takeover performance prediction**

**Generation of features and ground truth**

We used continuous decomposition analysis (CDA) to decompose the GSR signal into phasic and tonic components respectively [3]. After data pre-processing, we aggregated
the values of physiological data within a sliding 'time window' in order to fit time series data into the supervised learning framework [1]. The window size ranged from 1 to 40 seconds right before a TOR and the generated features were listed in Table 1. Model inputs included data on gaze behaviors, galvanic skin response, and heart rate, as well as traffic density and TOR lead time. In order to alleviate the individual differences, we normalized the feature values across participants via the min-max normalization approach. The subjective performance ratings were divided into two groups using the cut-off score median and were treated as the ground truth labels.

**Random Forest**

We used random forest models to predict drivers’ takeover performance for the following reasons. First, the relationships between drivers’ states, environments and takeover performance may be more complex than linear. Previous research has shown that random forest models generally outperform simple decision trees and boosted tree models [8]. Second, as an ensemble of binary decision trees for classification, random forests are robust for new data generalization and against training data overfitting [20]. Third, the random forest can produce feature importance, which increases the model interpretability.

We implemented two important techniques for model development via random forest algorithms: the bagging technique and a random sample of predictors for decision splits [6]. Bagging fits an algorithm on a set of bootstrapping samples, i.e., randomly selects many sets of samples with replacement from the training data. Classification in a bagging algorithm is based on a majority vote among the trees [5]. Each tree in the ensemble takes a random set of predictors at each node without replacement so that any two trees had weak or no correlations between them. In this research, we set this number as the square root of the number of predictors.

Among all the data, about 36.8 % of the total data forms the out-of-bag (OOB) samples. By permuting the out-of-bag data randomly across one predictor at a time and by measuring how much this permutation reduces the accu-

<table>
<thead>
<tr>
<th>Feature</th>
<th>Explanations</th>
</tr>
</thead>
<tbody>
<tr>
<td>HR indices</td>
<td>Mean, min, max and standard deviation of heart rate, inter-beat interval</td>
</tr>
<tr>
<td>GSR indices</td>
<td>Mean, max and standard deviation of GSR in phasic component</td>
</tr>
<tr>
<td>GSR peak fixation</td>
<td>The number of GSR peaks and peak rise time</td>
</tr>
<tr>
<td>Fixation fixation</td>
<td>Fixation number/duration in different area of interests (AOIs) (e.g., roads)</td>
</tr>
<tr>
<td>Saccade</td>
<td>Saccade number in AOIs (e.g., roads)</td>
</tr>
<tr>
<td>Pupil</td>
<td>Pupil diameter in AOIs (e.g., roads)</td>
</tr>
<tr>
<td>Blink</td>
<td>The number of blinks</td>
</tr>
<tr>
<td>Gaze dispersion</td>
<td>Standard deviation of the values for gaze angle (in radians) from eye point</td>
</tr>
<tr>
<td>Eyes-on-the-road</td>
<td>The proportion of time that participants’ gazes are on the road</td>
</tr>
<tr>
<td>Scan pattern</td>
<td>Eye transition patterns among different AOIs</td>
</tr>
<tr>
<td>Traffic density</td>
<td>Low or high oncoming traffic</td>
</tr>
<tr>
<td>TOR lead time</td>
<td>How early the vehicle issued TOR</td>
</tr>
</tbody>
</table>
racy of the model, we can get estimates of the feature importance in Matlab 2018b (MathWorks, MA). With 10-fold cross-validation, three sets of the hyper parameters were tuned for model training: the number of trees, minimum leaf size, and threshold for feature subsets.

In a binary classification problem, there are four possible outcomes: true positive (TP), false positive (FP), true negative (TN), and false negative (FN). In this paper, we used the traditional classification evaluation indicators Precision, Recall, Accuracy and F1-Score to carry out the evaluation of the experiment, which were defined as: Precision = \( \frac{TP}{TP + FP} \), Recall = \( \frac{TP}{TP + FN} \), Accuracy = \( \frac{TP + TN}{TP + FP + TN + FN} \), F1-score = \( \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \).

Results

Effect of Window Size

Using the median of the subjective ratings as the cut-off score, we classified takeover performance into two classes: good and bad. 34 physiological predictor variables and 2 environment variables were included as features in the random forest model. Due to system malfunctions, only 684 data points were left for model development. Generally, as shown in Figure 7, the predictive accuracy and F1-score increased while the time window increased.

The best model performance was achieved when the time window was 36 seconds. The best model accuracy and F1-score were 70.2% and 70.1% through 10-fold cross-validation (tuned hyper parameters: the number of trees - 200; minimum leaf size - 8; feature importance threshold - .18). Considering the difficulty in accessing a large time window in reality, a 18s time window was selected to predict drivers' subjective takeover performance in real-time. Figure 8 shows the confusion matrix with the 18s time window, and model accuracy and F1-score were both 67.4% (tuned hyper parameters: the number of trees - 250; minimum leaf size - 3; feature importance threshold - .18).

Effect of the subset of features

Figure 9 illustrates the out-of-bag estimates of feature importance of the 36 predictor variables when time window is 18s. Varying the threshold to include different sets of predictor variables, we found that the model had the highest accuracy and F1-score both at 67.4% when we only included the top 13 features with feature importance threshold at .18 (i.e. mean of heart rate and inter-beat-interval;
maximum of heart rate and inter-beat interval; standard deviation of inter-beat interval; maximum GSR; the number of GSR peaks and blinks; mean of pupil diameter in general and in NDRT specifically; vertical gaze dispersion; eyes-on-the-road percentage, and the probability that drivers transited eyes to driving scenes when they looked at NDRT).

However, if we included all sets of features, the model had accuracy and F1-score both at 64.9%. When feature importance threshold was 0.42, only vertical gaze dispersion, mean of heart rate and maximum inter-beat interval were included in the model. Then the model accuracy and F1-score were both 59.3%. Therefore, a subset of important features was able to generate better model performance.

**Discussion and conclusion**

This study develops a method to predict drivers’ takeover performance using their physiological data and environment parameters via random forest algorithm. The best predictive model can differentiate drivers’ takeover performance when they are engaged in the same type of NDRTs but with different levels of cognitive load with an accuracy of 70.2% and F1-score of 70.1%.

Future study can try a convolutional neural network (CNN) combined with long-short-term memory (LSTM) to predict drivers’ takeover performance since the physiological data is time-series. Moreover, objective driving behaviors can be explored to refine the ground truth in addition to subjective ratings. Instead of using dichotomous classification of takeover performance, we can increase the number of classes (e.g. bad, neutral, good; Or very bad, bad, neutral, good, very good) to see model prediction power.

The findings of the study can contribute to the design of an adaptive in-vehicle alert system. For example, if the system predicts that a driver is going to have bad takeover performance, a multi-modal display can be issued to help the driver realize the urgency of the event, increase situational awareness and allocate attention properly, resulting in better takeover performance. Eventually, it will enhance the safety and adoption of automated vehicles.

**Acknowledgements**

This project was supported by University of Michigan Mcity and in part by the National Science Foundation. The views expressed are those of the authors and do not reflect the official policy or position of State Farm®.

**REFERENCES**


Applications (AutomotiveUI ’19) (AutomotiveUI ’19).
ACM, New York, NY, USA, 70–90.


