Predicting Driver Takeover Performance in Conditionally Automated Driving

Na Du
Industrial and Operations Engineering, University of Michigan

Feng Zhou
Industrial and Manufacturing Systems Engineering, University of Michigan-Dearborn

Elizabeth M. Pulver
State Farm Mutual Automobile Insurance Company

Dawn M. Tilbury
Mechanical Engineering, University of Michigan

Lionel P. Robert
School of Information, University of Michigan

Anuj K. Pradhan
Industrial and Mechanical Engineering, University of Massachusetts Amherst

X. Jessie Yang
Industrial and Operations Engineering, University of Michigan

Accepted to be published in Accident Analysis & Prevention 08/22/2020

Manuscript type: *Research Article*

Running head: *Takeover performance prediction*

Corresponding author: X. Jessie Yang, 1205 Beal Avenue, Ann Arbor, MI 48015, Email: xijyang@umich.edu

Acknowledgment: This work 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®.

ABSTRACT

In conditionally automated driving, drivers have difficulty taking over control when requested. To address this challenge, we aimed to predict drivers' takeover performance before the issue of a takeover request (TOR) by analyzing drivers' physiological data and external environment data. We used data sets from two human-in-the-loop experiments, wherein drivers engaged in non-driving-related tasks (NDRTs) were requested to take over control from automated driving in various situations. Drivers' physiological data included heart rate indices, galvanic skin response indices, and eye-tracking metrics. Driving environment data included scenario type, traffic density, and TOR lead time. Drivers' takeover performance was categorized as good or bad according to their driving behaviors during the transition period and was treated as the ground truth. Using six machine learning methods, we found that the random forest classifier performed the best and was able to predict drivers' takeover performance when they were engaged in NDRTs with different levels of cognitive load. We recommended 3 s as the optimal time window to predict takeover performance using the random forest classifier, with an accuracy of 84.3% and an F1-score of 64.0%. Our findings have implications for the algorithm development of driver state detection and the design of adaptive in-vehicle alert systems in conditionally automated driving.

Keywords: Transition of control, predictive modeling, human-automation interaction, human-autonomy interaction, human-robot interaction.

1. Introduction

While automated vehicles are poised to revolutionize surface transportation, they introduce new challenges. One of the challenges is takeover transitions in conditionally automated driving (Ayoub, Zhou, Bao, & Yang, 2019; Zhou, Yang, & Zhang, 2020). In conditionally automated driving, drivers are no longer required to actively monitor the driving environment and are allowed to fully engage in non-driving-related tasks (NDRTs) (Society of Automotive Engineers, 2018). However, serving as a fallback for the automation, drivers are required to take over control of the vehicle whenever the automated system reaches its operational limit.

Previous studies showed that the limited driver-vehicle interaction in conditionally automated driving increases the difficulty for drivers to take over control when requested (Eriksson & Stanton, 2017; Gold, Körber, Lechner, & Bengler, 2016; Petersen, Robert, Yang, & Tilbury, 2019). In response to such difficulty, empirical studies have investigated the factors that influence drivers' takeover performance, including drivers' cognitive and emotional states (Du et al., 2020; Wan & Wu, 2018; Zeeb, Härtel, Buchner, & Schrauf, 2017) and driving environments (Gold et al., 2016; Li, Blythe, Guo, & Namdeo, 2018).

These studies shed light on the relationships between certain factors and takeover performance; for instance, high traffic density harmed takeover performance (Gold et al., 2016). However, with few exceptions (Braunagel, Rosenstiel, & Kasneci, 2017; Gold, Happee, & Bengler, 2018), little effort has been made to integrate these findings into computational models that are capable of predicting drivers' takeover performance in real time. In the present study, therefore, we aimed to fill the research gap and to predict drivers' takeover performance when they were engaged in NDRTs with different levels of cognitive load.

1.1 Factors influencing takeover performance

To facilitate takeover transitions, empirical research has been conducted to examine factors that influence drivers' takeover performance. The factors include drivers' cognitive and emotional states when performing different types of NDRTs (Du et al., 2020; Wan & Wu, 2018; Zeeb et al., 2017) in different driving environments (Gold et al., 2016; Li et al., 2018). Takeover performance consists of takeover timeliness (i.e., takeover reaction time) and takeover quality (e.g., speed, acceleration and jerk statistics, time/distance to collision statistics, lane deviation statistics, and crash rate).

The types of NDRTs have been found to influence takeover performance. Previous studies showed that compared with not performing an NDRT, those engaged in NDRTs had longer takeover reaction times, more crashes in high-traffic situations, and shorter minimum time to collision (TTC) (Eriksson & Stanton, 2017; Gold et al., 2016; Wan & Wu, 2018). The effects of NDRT modality on takeover performance were also explored. For example, Radlmayr, Gold, Lorenz, Farid, and Bengler (2014) and Wandtner, Schömig, and Schmidt (2018) reported that a visual task with handheld devices degraded takeover performance and led to a higher collision rate, while an auditory task led to comparable performance to a baseline without any task. Zeeb, Buchner, and Schrauf (2016) and Zeeb et al. (2017) explored the effects of manual and cognitive task load and found that a high level of manual task load increased reaction time and deteriorated takeover quality, while the effect of cognitive task load on takeover ability was dependent on the type of driver intervention. A high level of cognitive load lengthened the reaction time and deteriorated takeover quality in steering maneuvers but not braking maneuvers.

Driving environment factors include traffic density, road situations, and weather conditions. Heavy traffic density in takeover situations led to longer takeover time and worse takeover quality in the form of shorter time to collision, more collisions, and higher maximum accelerations (Gold et al., 2016; Körber, Gold, Lechner, & Bengler, 2016; Radlmayr et al., 2014). Li et al. (2018) showed that drivers' takeover reaction time to critical events in adverse weather conditions was longer on the highway compared to on city roads. Takeover request (TOR) lead time is the critical event onset for automation failures at the time of the TOR (McDonald et al., 2019). According to the complexity of driving environment and vehicle sensor capability, commonly used

TOR lead times range from 1 to 30 s (Eriksson et al., 2018). Research has demonstrated that shorter TOR lead time degraded takeover quality, as demonstrated by higher crash rates, greater maximum accelerations and greater standard deviation of steering wheel angle (Mok et al., 2015; van den Beukel & van der Voort, 2013; Wan & Wu, 2018).

Most of these studies focused on the effects of certain variables on takeover performance, providing valuable yet largely relational insights. For instance, heavy traffic density led to longer takeover time. However, knowing the relationships between certain factors and takeover performance is not enough to accurately predict a driver's takeover performance in the real world because many influential factors could interact with one another. Computational models capable of predicting drivers' takeover performance under various takeover conditions in real time are needed.

1.2 Predicting drivers' states through physiological measurements

With advances in wearable technology, it is possible to collect drivers' physiological signals, such as gaze behaviors, heart rate activity, and galvanic skin responses, for a reliable reflection of their cognitive and emotional states in conditionally automated driving.

Drivers' gaze behavior is a valid tool for measuring cognitive load (Gold et al., 2016; Luo et al., 2019; Solovey, Zec, Garcia Perez, Reimer, & Mehler, 2014; Wang, Reimer, Dobres, & Mehler, 2014; Zeeb et al., 2016) and visual scanning patterns have been shown to indicate situational awareness (Bertola & Balk, 2011; Ratwani, McCurry, & Trafton, 2010; Young, Salmon, & Cornelissen, 2013). For example, Gold et al. (2016) found that horizontal gaze dispersion was the most sensitive measure of drivers' cognitive demand in NDRTs during conditionally automated driving. Eyes-on-the-road percentage was found to be associated with drivers' situational awareness and attention capture of the driving environments (Molnar, 2017; Young et al., 2013).

Heart rate (HR) and heart rate variability (HRV) have both been used for assessing drivers' workload in real time (Mehler, Reimer, & Coughlin, 2012; Mehler, Reimer, Coughlin, & Dusek, 2009; Zhou, Alsaid, et al., 2020). Galvanic skin responses

(GSRs) were found to reflect drivers' mental activities, and their properties (amplitude, frequency) were used to indicate drivers' changes of arousal related to events (Collet, Clarion, Morel, Chapon, & Petit, 2009). GSRs have also been linked to drivers' workload and stress (Jones, Chapman, & Bailey, 2014; Schmidt, Decke, & Rasshofer, 2016; Wandtner et al., 2018).

Physiological data can thus be used to understand drivers' cognitive and emotional states by applying machine learning models to continuously monitored physiological data. The data captured via non-intrusive sensors can be used to build models that estimate drivers' states and their interactions with the driving environments. Drivers' physiological signals combined with environment factors are promising indicators to predict takeover performance in conditionally automated driving in real time (Braunagel et al., 2017).

1.3 Existing models for takeover performance prediction

Although a substantial amount of research has identified factors that influence drivers' takeover performance, there is a lack of research on the development of computational models for predicting drivers' takeover performance, with few exceptions (Braunagel et al., 2017; Gold et al., 2018).

To predict takeover performance, Gold et al. (2018) analyzed 753 takeover events using data from six driving simulator experiments and developed regression models. Their study modeled takeover performance measures (e.g., take-over time, minimum TTC, brake application and crash probability) as a function of the time-budget, traffic density, non-driving-related task, repetition, the current lane and driver's age. The models were validated using 729 takeover events from five additional experiments. The validation results showed that the regression models accurately predicted takeover time, time-to-collision and crash probability, and moderately predicted the brake application.

Braunagel et al. (2017) used machine learning algorithms to predict drivers' takeover quality (named as "takeover readiness" in the article). The study categorized takeover quality into low and high levels by analyzing driving parameters such as lane deviations. Data were collected from a driving simulator study with 81 participants. The first feature input was situation complexity with three levels decided by raters; the second set of features was the type of NDRTs performed by drivers; and the third set of features was drivers' gazes at the road. Using machine learning algorithms including k-nearest neighbors (kNN), support vector machine (SVM) with radial basis function (RBF) and linear kernel, Naive Bayes and linear discriminant, they predicted takeover quality with an accuracy of 79% and F1-score of 77%.

However, the above-mentioned models were developed and tested when drivers were engaged in different types of NDRTs (e.g., monitoring vs. reading), where apparent contextual cues existed to discriminate drivers' states. In daily life, even with a specific type of NDRTs such as writing an email, drivers' states can be rather different depending on the importance of the email. Also, some factors deliberately manipulated in the experiment settings such as emotions are not easily accessible in the real world. Although the advanced wearable technology has made it convenient to collect drivers' physiological signals to reflect their cognitive and emotional states, only gaze behaviors were used in previous studies.

1.4 The present study

Our study contributes to the literature in three aspects. First, our study aimed to predict drivers' takeover performance when they were engaged in a specific type of NDRTs with different levels of cognitive load. Second, in addition to gaze behaviors, we used drivers' heart rate indices and galvanic skin response indices to indicate their interaction with environments, which might improve prediction results. Third, our study employed a random forest model in addition to the machine learning models used in previous studies to predict takeover performance. Random forests have been proved to have great prediction performance for classification problems (Dietterich, 1997; McDonald, Lee, Schwarz, & Brown, 2014; Zhou, Alsaid, et al., 2020).

In this paper, data from two human subject experiments were used for model development. We collected drivers' galvanic skin responses (Collet et al., 2009; Mehler et al., 2012; Wintersberger, Riener, Schartmüller, Frison, & Weigl, 2018), heart rate activities (Bashiri & D Mann, 2014; Mehler et al., 2012), and gaze behaviors (Bertola & Balk, 2011; Radlmayr et al., 2014; Wang et al., 2014; Young et al., 2013), which have been used as valid signals to assess drivers' cognitive and emotional states and their situational awareness of the driving environments. Using drivers' physiological data and environment factors, we developed a random forest model that was able to predict drivers' takeover performance with an accuracy of 84.3% and an F1-score of 64.0% using a 3 s time window. Additionally, we identified the most important physiological measures for takeover performance prediction, which can be incorporated in practice to develop in-vehicle monitoring systems. Furthermore, the model can be used to guide the design of adaptive in-vehicle alert systems to improve takeover performance in conditionally automated driving.

2. DATASET

The data used in the development of algorithms were collected in two studies. Both studies complied with the American Psychological Association code of ethics and were approved by the institutional review board at the University of Michigan. The first study investigated the effects of cognitive load, traffic density, and TOR lead time on takeover performance. The second study examined the effects of scenario type and vehicle speed on takeover performance. Participants in both experiments wore the same set of physiological sensors. The similar experimental settings in both studies make it possible to combine the two datasets. At the same time, the varieties of takeover conditions from the two studies increase model generalizability.

2.1 Participants

A total number of 102 university students (mean age = 22.9; standard deviation [SD] = 3.8; range = 18–38; 40 females and 62 males) participated in Study 1 and 40 university students (mean age = 22.8, SD = 3.9; 20 females and 20 males) participated in Study 2. All of the participants had normal or corrected-to-normal vision and a valid driver's license. They received $30 in compensation for an hour of participation.

2.2 Apparatus and stimuli

Both studies were conducted in a fixed-base driving simulator from Realtime Technologies Inc. (RTI, MI, USA). The virtual world was projected on three front screens 16 ft away (120° field of view), one rear screen 12 ft away (40° field of view), and two side mirror displays (See Figure 1a).

This simulator was equipped with the Smart Eye four-camera eye-tracking system (Smart Eye, Sweden) that provided live head-pose, eye-blink, and gaze data (Figure 2a). The sampling rate of the eye-tracking system is 120 Hz. The Shimmer3 GSR+ unit (Shimmer, MA, USA) including GSR electrodes and photoplethysmogram (PPG) probe was used to collect GSR and HR data with a sampling rate of 128 Hz (Figure 2b). The iMotions software (iMotions, MA, USA) was used for physiological data synchronization and visualization in real time (Figure 2c).

The simulated vehicle was controlled by a steering wheel and pedal system embedded in a Nissan Versa car model. The vehicle was programmed to simulate SAE Level 3 automation, which handled the longitudinal and lateral control and navigation, and responded to traffic elements. Participants could press the button on the steering wheel to activate the automated mode, which was indicated by a green highlight on the dashboard and an auditory warning ("Automated mode engaged"). Once the AV reached its performance limit, an auditory TOR ("Takeover") would be issued with the green highlight turning to black background on the dashboard. Although the Level 3 automation is considered to continue functioning for a certain period of time after issuing the TOR (ISO, ISO/TR 21959-1:2020), we set the automated mode to be deactivated at the time of TORs for drivers to take over control of the vehicle.

The NDRT in both studies was a visual N-back memory task, adapted from the study of Jaeggi, Buschkuehl, Jonides, and Perrig (2008). The stimulus consisted of nine (3×3) squares with two human figures randomly in two of the nine squares. Each stimulus was presented for 500 ms in sequence with a 2,500–ms interval (Figure 3). 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. With different N values, participants were exposed to different cognitive load but the same manual and visual load. The reason for employing a visual task with manual input was that it simulated the eyes-off the road and hands-off the wheel condition. The task was running on an 11.6-in. touch screen tablet mounted in the vehicle (Figure 1b).

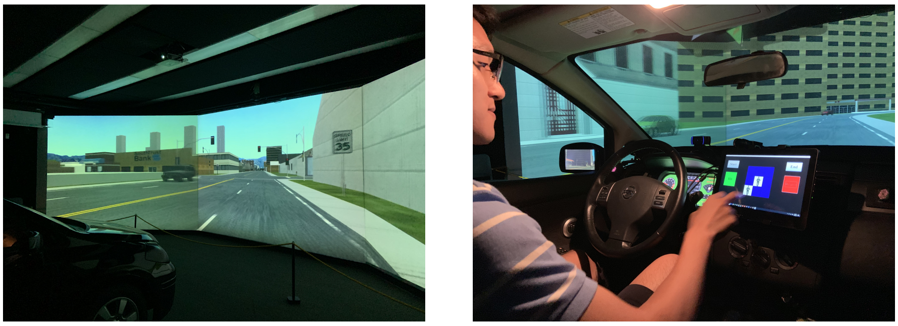

(a) (b)

Figure 1. RTI driving simulator at the UMTRI.

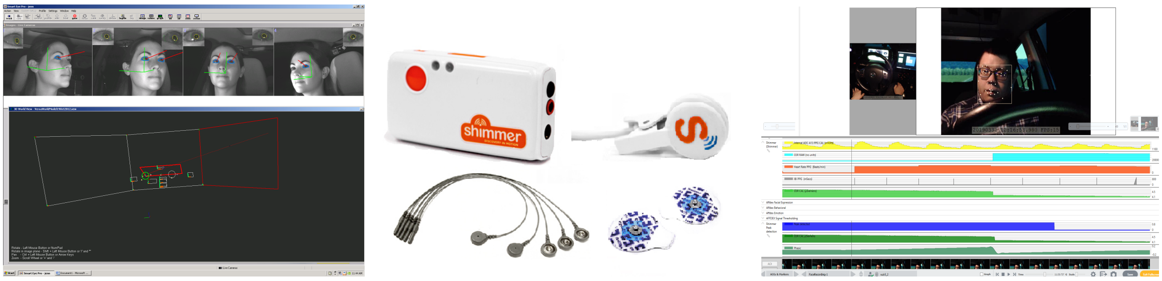

(a) (b) (c)

Figure 2. (a) Smart Eye. (b) Shimmer3 GSR+ unit. (c) iMotions software.

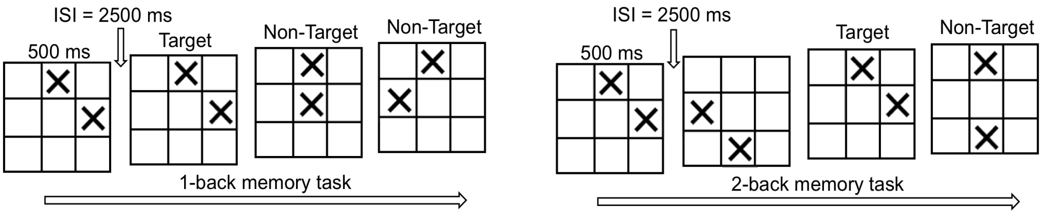


Figure 3. N-back memory task

2.3 Experimental design

Study 1 employed a within-subjects design with drivers' cognitive load, traffic density, and TOR lead time as independent variables. The cognitive load refers to driver cognitive load prior to TORs and was manipulated via the difficulty of the

NDRTs (low: 1-back memory task; high: 2-back memory task). The heavy– and no–traffic conditions had 15 and 0 oncoming vehicles per kilometer, respectively (Gold et al., 2016). The TOR lead time, which refers to the critical event onset for failures at the time of the TOR (McDonald et al., 2019), was set at 4 or 7 s (Eriksson & Stanton, 2017). Based on prior literature (Koo, Shin, Steinert, & Leifer, 2016; Miller et al., 2016; Molnar et al., 2018; Rezvani et al., 2016), 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 on the curve, (7) sensor error on the left curve, and (8) police vehicle on shoulder. The order of cognitive load, traffic density and TOR lead time was counterbalanced via an 8 × 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 Events 1 to 8, and the other half from Events 8 to 1.

Study 2 used a mixed design with scenario type (lane keeping vs. lane changing) as the between-subjects variable and vehicle speed (35 mph vs. 60 mph) as the within-subjects variable. Similar to the first study, eight scenarios were designed on the basis of realistic situations and previous literature (Koo et al., 2016; Miller et al., 2016; Naujoks, Mai, & Neukum, 2014; Rezvani et al., 2016; Zeeb et al., 2016). Lane-keeping scenarios, which required drivers to keep in the current lane, included (1) sensor error on the left curve, (2) construction zone on the left, (3) no lane markings on the curve, (4) sensor error on the right curve. Lane-changing scenarios, which required drivers to change to the neighboring lane, included (1) stranded vehicle ahead, (2) construction zone ahead, (3) construction barrier ahead, and (4) police vehicle on shoulder. According to the range of the Velodyne Lidar sensors (Velodyne Lidar, CA, USA), we set the distance between obstacle/entrance of the curve and the AV at 100 meters when the TOR was issued. Generally, traffic consisted of 15 oncoming vehicles per kilometer (Gold et al., 2016). The order of the vehicle speed was counterbalanced among participants. The order of scenarios was counterbalanced by having half of the participants drive from Events 1 to 4, and the other half from Events 4 to 1.

In both studies, drivers started from the right lane, and were asked to stay in the right lane before they engaged the automated mode. Thus, the AV was always in the right lane prior to the TORs and the objects could be pre-coded to appear in front of the vehicle in lane-changing scenarios. With two lanes in lane-changing scenarios, drivers could avoid the objects in their lane by changing to the adjacent lane because there were no other vehicles in the driver's direction. The speed of the subject vehicle was 35 mph in the urban/rural and 60 mph in the highway environments. The radius of curves was 400 meters in the highway and 100 meters in the urban/rural environments. Participants were asked to follow the speed limit throughout the drive.

2.4 Experimental Procedure

The procedures of the two studies were almost the same. After participants signed an informed consent form and completed an online demographics questionnaire, they were asked to track six targets on the front screen for eye-tracking calibration. Next, two GSR electrodes were attached to their left foot and the PPG probe to their left ear lobe. Participants were informed that there was no need to actively monitor the driving environments or take over control of the vehicle as long as the vehicle was in automated mode.

Participants had a 2-minute practice for the N-back memory task, followed by a 5-minute practice drive to get familiar with the simulator environment. Next, each participant drove two experimental drives (10–20 minutes each), each containing four (Study 1) or two (Study 2) 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. After about 90 s of NDRT, a TOR was issued unexpectedly, and participants were required to terminate the NDRT manually by pressing the "end" button on the tablet screen and take over the control immediately. When participants thought they had negotiated the takeover event, they were free to re-activate the AV mode. Participants were informed that they would get an additional $20 if their NDRT performance was ranked in the top 10 of all participants. The operation of the NDRT, the takeover, and the AV mode activation were repeated for each takeover event (Figure 4).

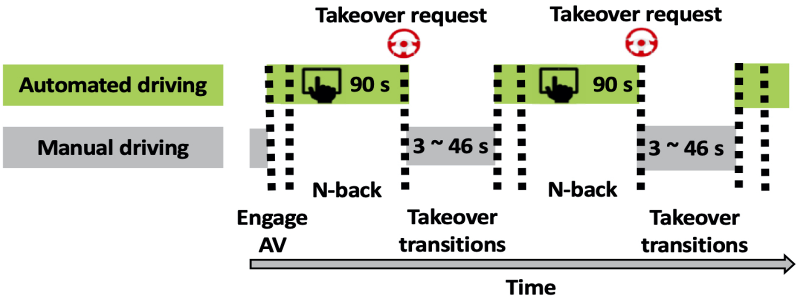


Figure 4. Illustration of the experimental procedure for two takeover events

3. TAKEOVER PERFORMANCE MODEL DEVELOPMENT

We collected drivers' physiological data, driving behaviors, and environment-related data. The physiological measures included heart rate indices, galvanic skin response indices and eye-tracking metrics. Because of malfunctions of the driving simulator and physiological sensors, data from 13 participants were excluded and those of the other 129 participants (i.e., 828 takeover scenarios) were available for further analysis.

To develop the prediction model, we first pre-processed the raw data and then extracted 37 features and set the ground truth. Next, we used a 10-fold nested cross-validation method to tune hyper-parameters, train models, and predict test instances for model comparisons. Particularly, we resampled the training dataset and normalized the entire dataset before performing the classification. Figure 5 shows the modeling process.

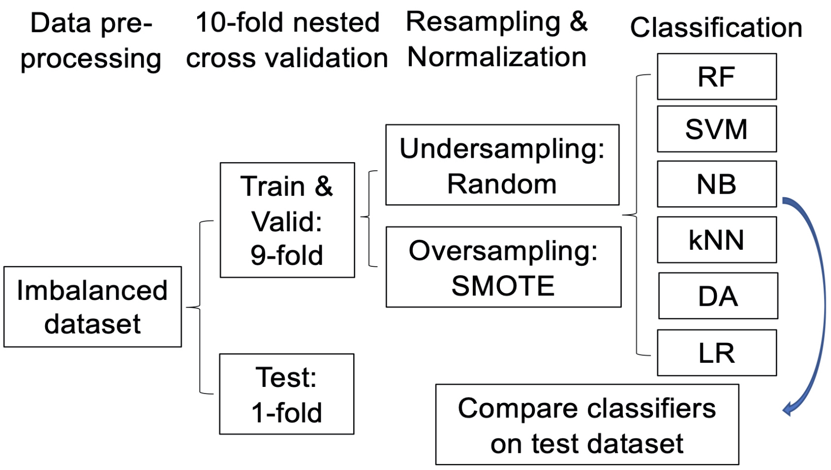


Figure 5. Modeling process (RF = random forest; SVM = support vector machine; NB = Naive Bayes; kNN = k-nearest neighbors; DA = discriminant analysis; LR = logistic regression).

3.1 Data pre-processing

For GSR signals, we used continuous decomposition analysis (CDA) to decompose the GSR signal into phasic and tonic components, respectively, via Ledalab in Matlab (Benedek & Kaernbach, 2010). Then we used the phasic component for further feature extraction because it is responsible for relatively rapid changes in response to specific events in the GSR signal (order of seconds). Heart rate measures were extracted from the raw RR interval using iMotions software. For eye-tracking data, only data points with high gaze quality value (threshold recommended by Smart Eye: .5) were recorded and used for analysis.

3.2 Feature generation and ground truth

To fit time series data into the supervised learning framework, we aggregated the values of physiological data within a sliding "time window" and calculated various statistics (Anderson, 2011). The end of the time window is the time of a TOR, and the start of the time window is X seconds before the TOR, ranging from 1 to 30 s. Model inputs included data on gaze behaviors, galvanic skin response indices, and heart rate indices, as well as environment factors. The generated features are listed in Table 1. A fixation is defined as "a relatively stable eye-in-head position within some threshold of dispersion (typically $\sim 2°$) over some minimum duration (typically 100-200 ms), and with a velocity below some threshold (typically 15-100° per second)" (Jacob & Karn, 2003). In the Smart Eye eye-tracking system, all frames with a gaze velocity below the fixation threshold (100° per second) were treated as a fixation. All frames with the gaze velocity above the saccade threshold (100° per second) were treated as a saccade. We categorized area of interests (AOIs) into driving scenes, the NDRT tablet, and other areas. The number and average duration of fixations and saccades were accumulated within the certain AOI. The scan pattern is the probability of eyes switching from one AOI to another. Traffic density, TOR lead time, and scenario type were used to describe the driving environments because they indicated the predictability, criticality, and urgency of the takeover scenarios (Gold, Naujoks, Radlmayr, Bellem, & Jarosch, 2017). To reduce the potential impact of individual differences, we normalized the feature values across participants using the min-max normalization approach.

We used driving behaviors during takeover transitions to assess drivers' takeover performance. As shown in Table 2, for different takeover scenarios, we selected different metrics in the assessment. Minimum TTC was calculated only for the lane-changing scenarios, and standard deviation of road offset was calculated only for the lane-keeping scenarios. All the driving variables were calculated following prior studies (Clark & Feng, 2017; Du et al., 2020). If any of the calculated TOR reaction time, maximum resulting acceleration, and standard deviation of road offset values were larger than $\mu + 2\sigma$, we categorized a takeover transition as a bad performance. For minimum TTC, because the value of $\mu - 2\sigma$ was negative, we performed a log transformation first and categorized a takeover transition as bad if log(minimum TTC) was lower than $\mu - 2\sigma$ (Braunagel et al., 2017). For a particular takeover event, as long as one of the driving variables in a certain takeover scenario was categorized as a bad performance, we labeled the scenario as a bad takeover performance. Scenarios that led to collisions were also categorized as bad performances. Eventually, we got an imbalanced dataset with 109 "bad performance" labels and 719 "good performance" labels. The reasons that we used categorical takeover performance rather than individual driving variables as model output were that (1) it combines multiple aspects of driving behaviors and (2) it is easy to be explained to drivers and more practical to guide driver behaviors.

TABLE 1: *Descriptions of generated features (HR = heart rate; min = minimum; max = maximum; GSR = galvanic skin responses; NDRT = non-driving-related task; TOR = takeover request).*

Feature	Explanations
HR indices	Mean, min, max, and standard deviation of heart rate, inter-beat interval
GSR indices	Mean, max, and standard deviation of GSR in phasic component
GSR peak	The number of GSR peaks, and peak rise time
Fixation	Fixation number and duration in different areas of interests (AOIs) (i.e., driving scenes and NDRT tablet)
Saccade	Saccade number in different AOIs (i.e., driving scenes and NDRT tablet)
Pupil	The mean and standard deviation of pupil diameter in different AOIs (i.e., driving scenes and NDRT tablet)
Blink	The number of blinks
Gaze dispersion	Standard deviation of the values for gaze angle from right front (radians)
Eyes-on-the-road	The proportion of time that participants' gazes are on the road
Scan pattern	The probability of eyes switching from one AOI to another (i.e., the probability that drivers transited eyes from driving scenes to NDRT tablet, from NDRT tablet to driving scenes, or from other areas to driving scenes)
Traffic density	No or heavy oncoming traffic
Scenario type	Lane-keeping or lane-changing scenarios
TOR lead time	Short (3-4s) or long (6-7s) TOR lead time

TABLE 2: *Takeover situations and corresponding driving behavior variables to determine takeover performance (TOR = takeover request; min = minimum; max = maximum; TTC = time to collision).*

Takeover reactions	Driving behavior variables (range for bad performance group)		
Lane changing	TOR reaction time ($> \mu + 2\sigma$)	Max resulting acceleration ($> \mu + 2\sigma$)	log(Min TTC) ($< \mu - 2\sigma$)
Lane keeping	TOR reaction time ($> \mu + 2\sigma$)	Max resulting acceleration ($> \mu + 2\sigma$)	Standard deviation of road offset ($> \mu + 2\sigma$)

3.3 Model development

The takeover performance prediction model was trained with a random forest model considering the following justifications. First, as an ensemble method, random forests are robust for new data generalization and against training data overfitting (Quinlan et al., 1996). Second, random forests can give us feature importance and makes models interpretable. Five other machine learning approaches mentioned in prior literature were applied for comparisons: k-nearest neighbors (kNN), support vector machine (SVM), Naive Bayes (NB), discriminant analysis (DA), and logistic regression (LR).

Considering the challenge of human behavior data collection, we used a 10-fold nested cross-validation method to train models and compare test results (J. J. Lee, Knox, Baumann, Breazeal, & DeSteno, 2013; Varma & Simon, 2006). As shown in Figure 5, the 9-fold training and validation set (N = 116 subjects) was used to tune the hyper-parameters with the inner loop and then create classifiers. To handle the imbalanced dataset during the training, we employed a hybrid method of undersampling and oversampling (Choirunnisa & Lianto, 2018). The elimination process was done by deleting 300 good takeover performance scenarios randomly (Prusa, Khoshgoftaar, Dittman, & Napolitano, 2015). Then we used Synthetic Minority Over-sampling Technique (SMOTE) to create a balanced training and validation dataset with 678 data points (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). Table 3 demonstrates the training procedures of six machine learning approaches. The model assessment was based on the remaining 1-fold testing set (N = 13 subjects) with the outer loop. Notably, the subject data used for testing were not seen in the model training and validation stage. The random selection of 1-fold test dataset assumed that its distribution of good and bad takeover performance scenarios was similar to the whole dataset. With a 10-fold cross-validation, we can make sure all the data points in the dataset would appear once in the test dataset. The training and evaluation of the algorithm were implemented in Matlab 2018b (MathWorks, MA, USA).

TABLE 3: *Machine learning techniques and training process*

Machine learning approach	Techniques	Hyper-parameters
Support vector machine (SVM)	Embed the data in another dimensional space and find a soft margin that separates the classes with minimum classification error (Chen, Wu, Ying, & Zhou, 2004)	Kernel, Regularization parameter
Naive Bayesian (NB)	Use maximum likelihood estimation to estimate parameters (i.e., prior probability and likelihood) (Rish et al., 2001)	None
Random forest (RF)	Fit an algorithm on a set of bootstrapping samples (bagging) and predictors, i.e., randomly select training samples with replacement and take a random set of predictors at each node without replacement. Repeat many times to form an ensemble of trees (Breiman, 1996, 2001)	Tree number, Predictor number per split, Leaf size
k-nearest neighbor (kNN)	Calculate Euclidean distance between labeled and unlabeled points to find the k-nearest neighbors. Use the majority vote criteria to decide unlabeled points (Keller, Gray, & Givens, 1985)	k
Discriminant analysis (DA)	Find separating hyperplane using parameter estimation (Friedman, 1989)	Discriminant type, Regularization parameter
Logistic regression (LR)	Estimate the parameters of a logistic model (S.-I. Lee, Lee, Abbeel, & Ng, 2006)	Regularization parameter

3.4 Model evaluation

In a binary classification problem, there are four possible outcomes: true positive (TP), false positive (FP), true negative (TN), and false negative (FN). TP is the number of positive samples predicted as a positive class, FP is the number of negative samples predicted as a positive class, FN is the number of positive class samples predicted as a negative class and TN is the number of negative samples predicted as negative class. In this paper, we used four classification evaluation indicators, including Precision, Recall, Accuracy, and F1-score, to carry out the evaluation of the model performance, which were defined as:

$$Precision = \frac{TP}{TP\ +\ FP} \tag{1}$$

$$Recall = \frac{TP}{TP\ +\ FN} \tag{2}$$

$$Accuracy = \frac{TP\ +\ TN}{TP\ +\ FP\ + TN\ +\ FN} \tag{3}$$

$$F1 - score = \frac{2 \times Precision\ \times Recall}{Precision\ +\ Recall} \tag{4}$$

Precision manifests how well the model predicts (i.e., a measure of exactness) and recall manifests how well the model does not miss the target (i.e., a measure of completeness). The F1 measure is the weighted harmonic mean of the two and represents a realistic measure of model performance.

The receiver operating characteristic (ROC) curve plots the true positive rate (TPR) against the false positive rate (FPR) at different thresholds (i.e., classifier boundary). The area under the curve (AUC) ranges from 0 to 1 and represents the degree of separability. A higher value of AUC indicates better model performance. When AUC is 0.5, it means the model does not have any class separation capability.

4. RESULTS

To improve the robustness of machine learning results, we ran the 10-fold cross-validation 30 times (i.e., 30 different random seeds) for every machine learning method at each time window. We first ran an omnibus analysis of variance (ANOVA) to compare the performance of the six machine learning methods. After that, we compared the random forest model with the other five methods using the pairwise *t*-test to see whether the random forest model had the best performance. Similarly, we compared the prediction results of the random forest model with different feature subsets against the full feature model using pairwise *t*-test. We examined the effects of time window and individual feature on random forest prediction performance using ANOVA. All post hoc comparisons used a Bonferroni α correction.

4.1 Model performance comparisons

Figures 6 and 7 show the average model accuracy and F1-score at different time windows. There was a main effect of machine learning approaches on the prediction accuracy ($F(5, 5399) = 13550, p < .001$) and F1-score ($F(5, 5399) = 4705, p < .001$). Table 4 shows the pairwise t-tests comparing the predictive performance of the random forest model with the other five models across different time windows. The results indicate that our proposed random forest model outperformed the other five models across time windows.

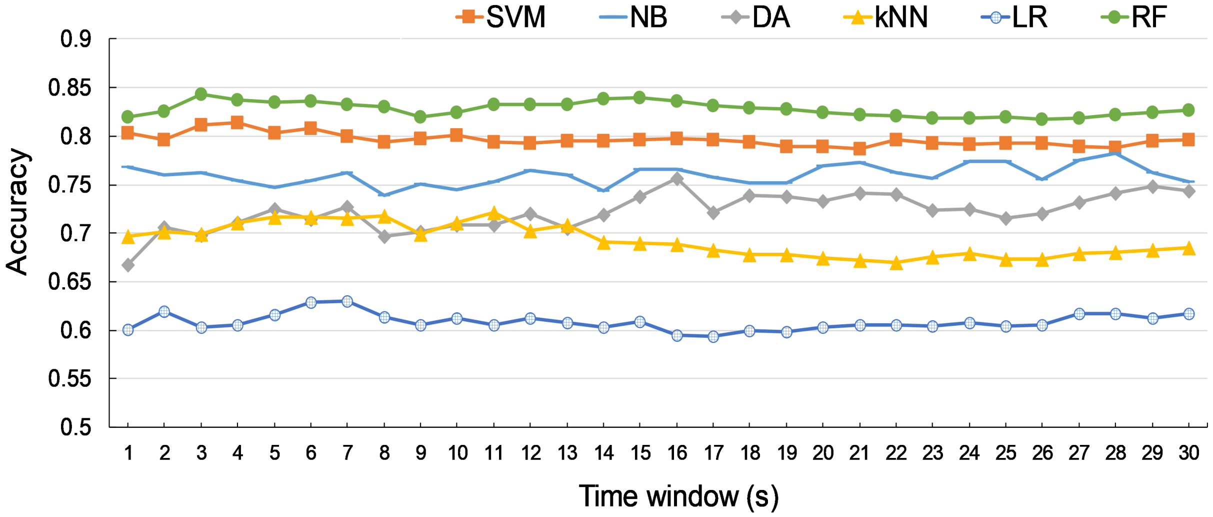


Figure 6. Prediction accuracy of six machine learning approaches under different time windows (SVM = support vector machine; NB = Naive Bayes; DA = discriminant analysis; kNN = k-nearest neighbors; LR = logistic regression; RF = random forest).

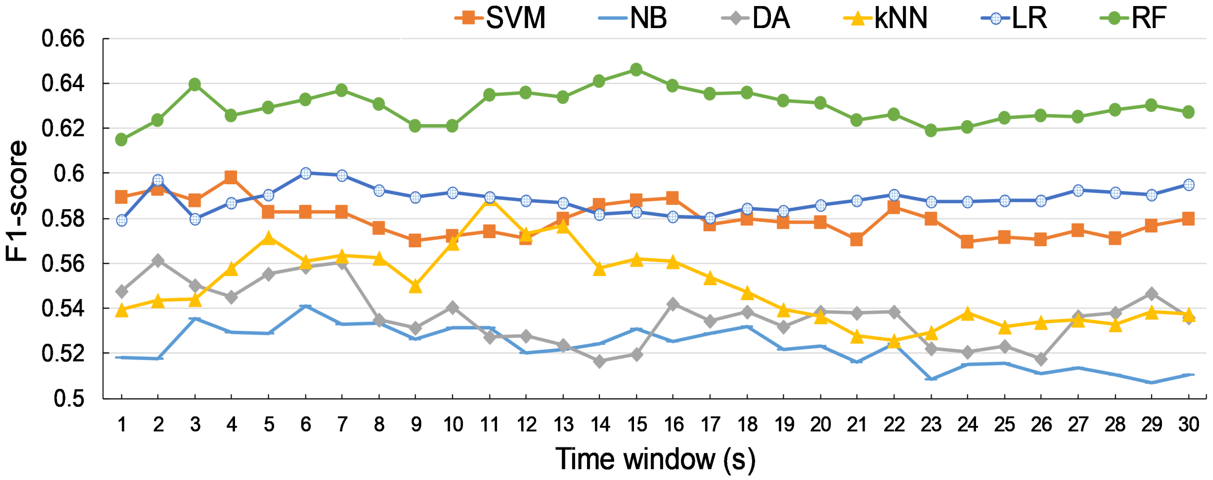


Figure 7. F1 scores of six machine learning approaches under different time windows (SVM = support vector machine; NB = Naive Bayes; DA = discriminant analysis; kNN = k-nearest neighbors; LR = logistic regression; RF = random forest).

TABLE 4: *The mean prediction accuracy and F1-score of machine learning approaches across time windows and their comparisons to the random forest model.*

Algorithm	Accuracy				F1-score			
	mean	SD	*t*-test statistic	p-value	mean	SD	*t*-test statistic	p-value
Random forest	.828	.012	-	-	.630	.015	-	-
Support vector machine	.796	.013	60.5	p<.001	.580	.019	72.4	p<.001
Naive Bayes	.760	.033	49.0	p<.001	.523	.022	107	p<.001
Discriminant analysis	.722	.021	134	p<.001	.537	.017	131	p<.001
k-nearest neighbor	.692	.020	209	p<.001	.550	.020	111	p<.001
Logistic regression	.609	.016	342	p<.001	.588	.009	74.5	p<.001

Figure 8 shows the ROC curves of the random forest and the other five machine learning approaches with the optimal hyper-parameters. The curve of the random forest is above and to the left of the other five curves at the majority of thresholds. Consistent with accuracy and F1-score results, the ROC curve comparisons demonstrated that the random forest model outperformed the other five models.

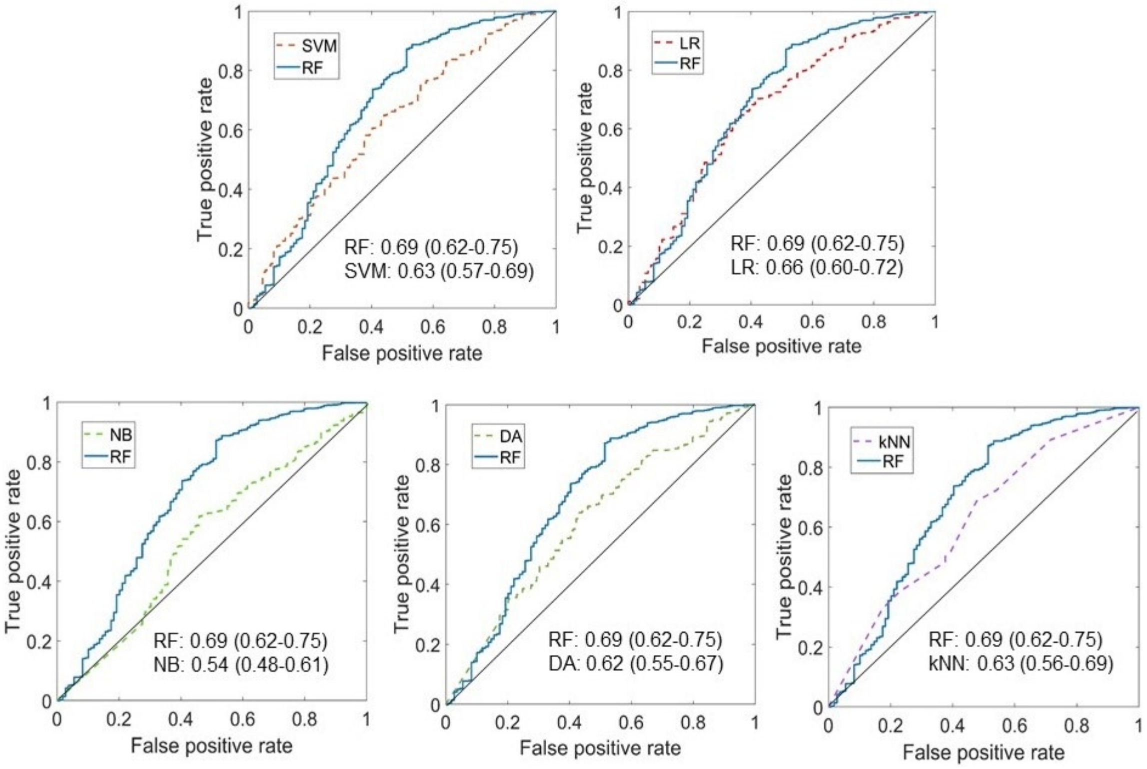


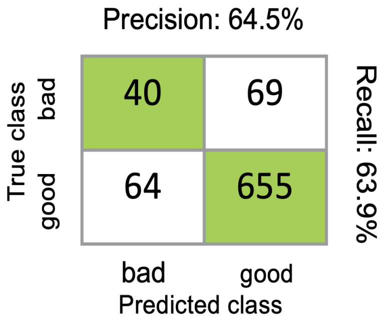
Figure 8. Receiver operating characteristic comparison plots for the random forest (RF) model and the five other models (SVM = support vector machine; LR = logistic regression; NB = Naive Bayes; DA = discriminant analysis; kNN = k-nearest neighbors). The bootstrapped (#1,000) confidence intervals are indicated within the parentheses.

4.2 Effects of window size on random forest prediction results

There was a main effect of time window on the random forest model accuracy ($F(29,899) = 16, p < .001$) and F1-score ($F(29,899) = 9, p < .001$). When applying an algorithm in real-world driving, a time window with shorter size and better prediction performance is preferred. According to Figures 6 and 7, we recommend 3 s as the optimal time window to predict takeover performance, with an average F1-score of 64.0% and accuracy of 84.3% (tuned hyper-parameters: the number of trees = 300; minimum leaf size = 2; the number of predictors per decision split = 6). Post hoc analysis showed that F1-score at the 3 s time window significantly outperformed the rest of the time windows except 5-8 s, 11-20 s, and 28-30 s (see Figure 7). Accuracy at the 3 s time window significantly outperformed the rest of the time windows except 4 s, 6 s, 11 s, and 13-16 s (see Figure 6).

4.3 The confusion matrix and feature importance

Figure 9 shows the confusion matrix when the time window was 3 s. The precision was 64.5% and the recall was 63.9%, accounting for balanced completeness and exactness of prediction.



	bad	good
bad	40	69
good	64	655

Figure 9. Confusion matrix when time window was 3s

Furthermore, by permuting the out-of-bag data (i.e., 36.8% of the total data that were not in the bootstrap samples) randomly across one predictor at a time and by measuring how much this permutation reduced the accuracy of the model, we estimated the feature importance. The values indicate each feature's relative importance in predicting the takeover performance (the larger values are, the more important features are). Figure 10 illustrates the out-of-bag estimates of feature importance of the 37 predictor variables when the time window was 3 s. Table 5 lists the top 16 important predictor variables. As shown in the table, we found that some heart rate indices and GSR indices (e.g., maximum and mean phasic GSRs, mean of heart rate) were important in predicting takeover performance, but were not included in prior takeover performance algorithm development (Braunagel et al., 2017; Gold et al., 2018).

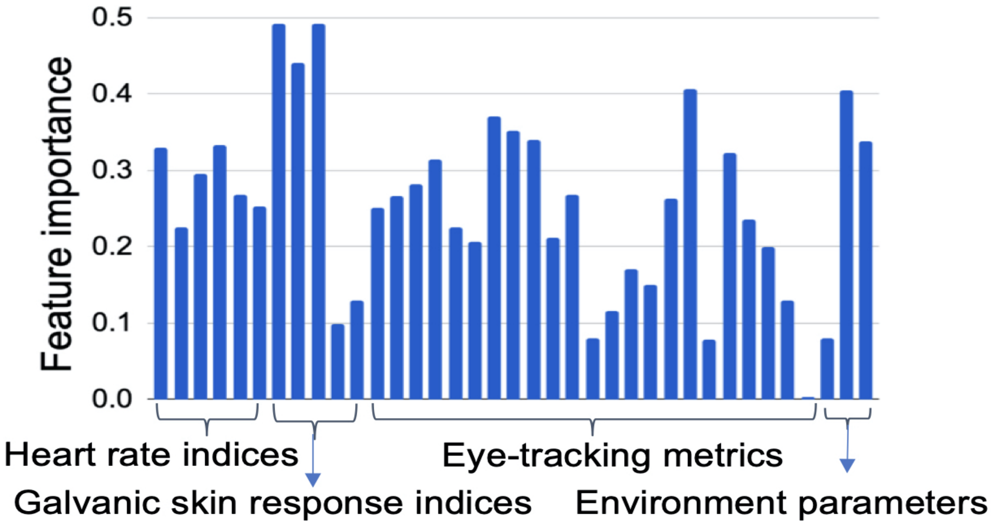


Figure 10. Feature importance when time window was 3s

TABLE 5: *The top 16 important features when time window was 3s (GSR = galvanic skin response; NDRT = non-driving-related task).*

Feature descriptions	Importance
Maximum of GSR in phasic component	.492
Mean of GSR in phasic component	.491
Standard deviation of GSR in phasic component	.441
Vertical gaze dispersion	.406
Scenario type	.404
Fixation duration	.371
Fixation duration on the driving scene	.352
Fixation duration on the NDRT	.341
Takeover lead time	.338
Mean of inter-beat interval	.333
Mean of heart rate	.330
Eyes-on-the-road percentage	.323
Saccade number on the driving scenes	.314
Maximum heart rate	.295
Fixation number on the driving scenes	.282
Standard deviation of inter-beat interval	.268

4.4 Effects of features on random forest prediction results

The main effect of feature set on the model accuracy ($F(3, 119) = 304, p < .001$) and the F1-score ($F(3, 119) = 146, p < .001$) were significant at the 3 s time window. We found that the accuracy and F1-score of the random forest model using the full feature set were significantly higher than the accuracy and F1-score using other combinations of feature subsets at the 3 s time window (Figure 11 and Table 6). To be specific, if only environment factors were used as the features, the average prediction accuracy and F1-score were only .758 and .611, respectively. If only physiological data were used as features, the average prediction accuracy was .770 and F1 score was 0.563. This suggests that a combination of environment features and features indicating drivers' states are necessary to build a model with high performance. The model using environment factors and eye-tracking metrics as features had an average accuracy of 0.818 and F1-score of 0.615 at the 3 s time window. After adding heart rate and galvanic skin response indices as features, the average model accuracy increased to 0.843 and F1-score increased to 0.640.

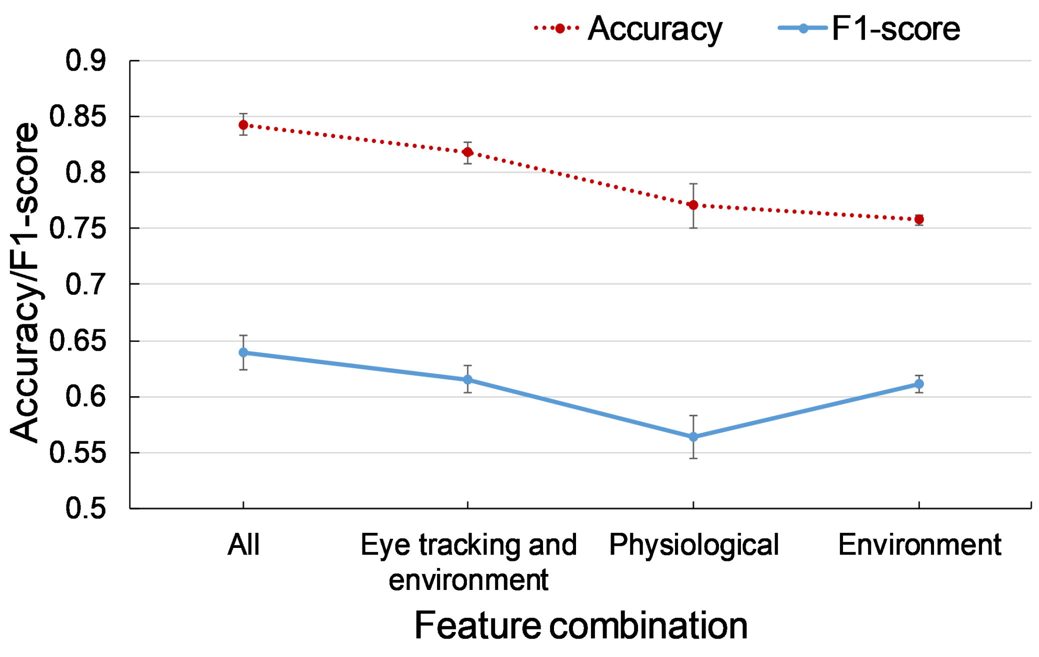


Figure 11. Prediction accuracy and F1-score of random forests with different feature subsets at the 3 s time window. Error bar indicates 1 standard deviation.

TABLE 6: *Random forest prediction accuracy and F1-score with different feature subsets at the 3 s time window and their comparisons to the full feature model.*

Feature subsets	Accuracy				F1-score			
	mean	SD	*t*-test statistic	p-value	mean	SD	*t*-test statistic	p-value
All	.843	.010	-	-	.640	.015	-	-
Eye-tracking and environment	.818	.010	11.2	p<.001	.615	.013	10.9	p<.001
Physiological	.770	.020	17.2	p<.001	.563	.019	19.7	p<.001
Environment	.758	.005	42.7	p<.001	.611	.008	8.82	p<.001

In addition, we ordered features according to the average feature importance values. Next, we built a random forest model with the most important feature, and then added features with lower importance one by one to build another 36 models. As shown in Figure 12, the model accuracy and F1-score generally increased at the beginning when more features were added but reached a plateau when 16 or more features were included in the model. There was a main effect of feature numbers on the model accuracy ($F(36, 1109) = 3718, p < .001$) and F1-score ($F(36, 1109) = 293, p < .001$). Post hoc analysis showed that the F1-score of the full feature model was significantly higher than that for models with fewer than the top 9 important features, and accuracy of the full feature model was significantly higher than that of the models with fewer than the top 16 important features.

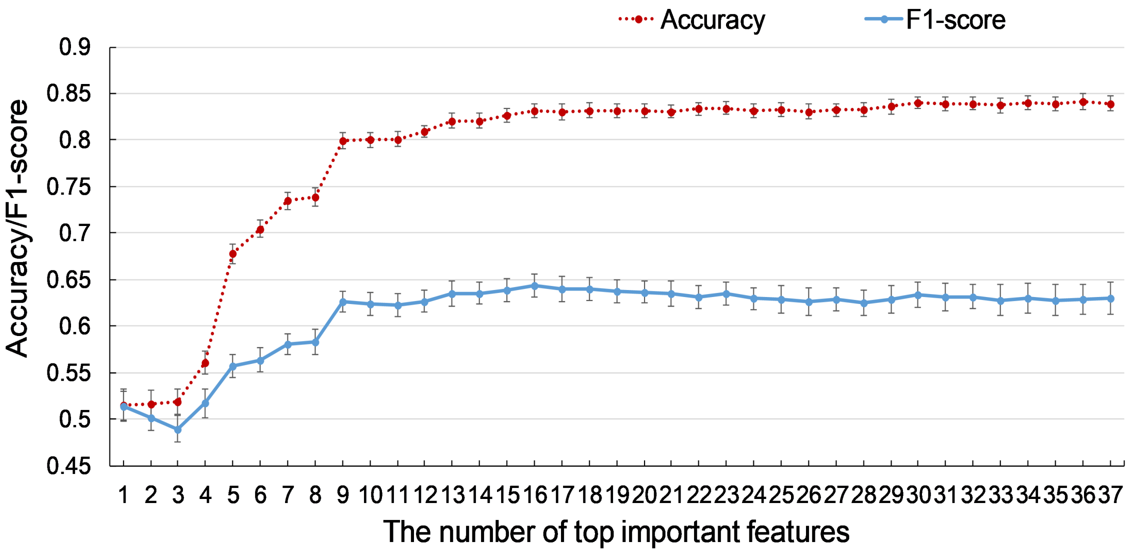


Figure 12. Model accuracy and F1-score with different numbers of top important features. Error bar indicates 1 standard deviation.

5. DISCUSSION

5.1 Model performance comparisons

Our study compared the random forest model with the other five machine learning approaches used in prior literature for takeover performance prediction. As indicated by the results of model accuracy, F1-score, and ROC curve comparisons, the random forest approach outperformed the other classification approaches. Consistent with previous studies on drivers' fatigue and drowsiness detection (McDonald et al., 2014; Zhou, Alsaid, et al., 2020), the random forest approach also showed its supremacy for takeover performance prediction. It might be because random forests aggregate the results of many bootstrap aggregated (bagged) decision trees, which reduces the effects of overfitting and improves generalization.

5.2 Effects of window size on random forest prediction results

As the random forest outperformed other machine learning approaches, we examined the prediction performance of random forests under different time window sizes. The results showed that the window size significantly influenced random forest prediction performance. However, such a relationship was not linear. One of the explanations could be that we used a mixture of physiological signals as model inputs. Some physiological signals (e.g., pupil diameter) perform better with a shorter window size because they change rapidly according to the changes in the driver's cognitive workload (Kramer et al., 2013). Some physiological signals (e.g., heart rate) perform better with a longer window size because it can provide an overall understanding of the driver's mental state (Solovey et al., 2014). Future research is needed to explore model performance with customized time windows for different physiological signals.

It was important to find an optimal window size to calculate physiological features for model development in this study. Considering the implementation in real-world driving, a time window with shorter size and better prediction performance is preferred. Thus, we recommend 3 s as the optimal time window to predict takeover performance, with an accuracy of 84.3% and an F1-score of 64.0%. The post hoc analysis showed that the selection of time window for such performance is not unique. Time windows with a size of 6 s, 11 s, and 13-16 s led to similar prediction performance. Although the exact time window might be slightly different in the real world given the differences of situational and behavioral parameters, our study provides important insights on window size recommendation for the development of driver state detection systems.

Different from previous studies, our model has a finer granularity and can predict drivers' takeover performance when they are engaged in a specific type of NDRTs with different levels of cognitive load. Such application differences make it infeasible to compare the exact accuracy and F1-score values with those in previous models. Because the test cases in our model prediction are from different participants and are not seen in the training set, our model can be used to predict takeover performance of a new driver who does not have historical data.

5.3 Effects of features on random forest prediction results

Drivers' galvanic skin responses, heart rate activities, and eye movements with a combination of environment factors were used to predict drivers' takeover performance. Compared to Braunagel et al. (2017), we added GSR indices and HR indices for model development. Our results showed an improvement of model performance with a full set of features compared to other feature subsets (i.e., physiological data only, environment data only, eye-tracking and environment data). This aligns with the previous studies because all these physiological signals reflected drivers' states and interactions with driving environments (Bertola & Balk, 2011; Mehler et al., 2012; Radlmayr et al., 2014; Ratwani et al., 2010; Wang et al., 2014; Young et al., 2013).

Furthermore, we identified the most important features (e.g., maximum phasic GSR, gaze dispersion, scenario type, and mean of inter-beat interval) for model development. Although the model performance increased at the beginning as more features were added, it reached a plateau when 16 or more features were included. With the top 16 important features, we were able to develop a random forest model with comparable performance to the full feature model. Notably, the top 16 important features were extracted from galvanic skin responses, heart rate activities, eye movements, and environment factors, demonstrating the importance of all these data sources. Utilizing the advances of wearable technology and vehicle sensors, these features can be collected in a minimally invasive manner to predict drivers' takeover performance in real time.

5.4 Limitations and future work

Several limitations should be taken into consideration in the future. First, this study used a snapshot of the time-series data as model inputs without considering the complexity of sequence dependence among the data. Future study could try a convolutional neural network (CNN) combined with long-short-term memory (LSTM) to predict drivers' takeover performance using a larger dataset. Second, the ground truth was determined by drivers' driving behaviors. It is necessary to propose a standard set of metrics for measuring takeover performance. An ensemble method combining subjective ratings, driving behaviors and video coding can be explored to provide a more robust ground truth label of takeover performance. Third, instead of using dichotomous classification of takeover performance, we could increase the number of classes (e.g., bad, neutral, good; or very bad, bad, neutral, good, very good) or use regression to see model prediction power. Fourth, this study only recruited young adult participants with few AV experiences and each participant only experienced four or eight takeover scenarios in the whole experiment. Future studies could recruit participants from different ages, AV experience levels, and training groups. Then the individual characteristics and power law of learning could be taken into account as model inputs to increase the generalization of models (Forster et al., 2019).

5.5 Implications

Our study is a preliminary effort to predict drivers' takeover performance for designing advanced driver monitoring systems. With the advances of technologies in connected automated vehicle systems, real-time road environments such as traffic situations can be accessed easily in the future. Predictive model performance can be improved when data from various drivers engaging in different NDRTs in diverse environments are available for model training. The model outputs can contribute to the design of adaptive in-vehicle alert systems in conditionally automated driving. Specifically, if the system predicted that a driver would not be able to take over control successfully, a multi-modal display could be designed to help the driver realize the urgency of the event, augment situational awareness and allocate attention properly. Eventually, it could improve drivers' takeover performance and enhance the safety and adoption of automated vehicles.

6. CONCLUSION

This study developed a random forest model to predict drivers' takeover performance in conditionally automated driving. In contrast to previous models capable of predicting drivers' takeover performance when they performed different types of NDRTs, our model has a finer granularity and is able to predict takeover performance when drivers are engaged in a specific type of NDRTs. The results showed that the random forest classifier has an accuracy of 84.3% and an F1-score of 64.0% using a 3s time window, which outperformed other machine learning models used in prior studies. In addition, we identified the most important physiological measures for takeover performance prediction, and they can be used for developing in-vehicle monitoring systems. Such models can be used to guide the design of adaptive in-vehicle alert systems to improve takeover performance in conditionally automated driving in the future.

References

Anderson, T. W. (2011). *The statistical analysis of time series* (Vol. 19). John Wiley & Sons.

Ayoub, J., Zhou, F., Bao, S., & Yang, X. J. (2019). From manual driving to automated driving: A review of 10 years of autoui. In *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '19)* (pp. 70–90). New York, NY, USA: ACM.

Bashiri, B., & D Mann, D. (2014). Heart rate variability in response to task automation in agricultural semi-autonomous vehicles. *The Ergonomics Open Journal*, *7*(1), 6–12.

Benedek, M., & Kaernbach, C. (2010). A continuous measure of phasic electrodermal activity. *Journal of neuroscience methods*, *190*(1), 80–91.

Bertola, M. A., & Balk, S. A. (2011). *Eyes on the road: A methodology for analyzing complex eye tracking data.*

Braunagel, C., Rosenstiel, W., & Kasneci, E. (2017). Ready for take-over? a new driver assistance system for an automated classification of driver take-over readiness. *IEEE Intelligent Transportation Systems Magazine*, *9*(4), 10–22.

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123–140.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321–357.

Chen, D.-R., Wu, Q., Ying, Y., & Zhou, D.-X. (2004). Support vector machine soft margin classifiers: error analysis. *Journal of Machine Learning Research*, *5*(Sep), 1143–1175.

Choirunnisa, S., & Lianto, J. (2018). Hybrid method of undersampling and oversampling for handling imbalanced data. In *2018 international seminar on research of information technology and intelligent systems (isriti)* (pp. 276–280).

Clark, H., & Feng, J. (2017). Age differences in the takeover of vehicle control and engagement in non-driving-related activities in simulated driving with conditional automation. *Accident Analysis & Prevention*, *106*, 468–479.

Collet, C., Clarion, A., Morel, M., Chapon, A., & Petit, C. (2009). Physiological and behavioural changes associated to the management of secondary tasks while driving. *Applied ergonomics*, *40*(6), 1041–1046.

Dietterich, T. G. (1997). Machine-learning research. *AI magazine*, *18*(4), 97–97.

Du, N., Zhou, F., Pulver, E., Tilbury, D. M., Robert, L. P., Pradhan, A. K., & Yang, X. J. (2020). Examining the effects of emotional valence and arousal on takeover performance in conditionally automated driving. *Transportation research part C: emerging technologies*, *112*, 78–87.

Eriksson, A., Petermeijer, S. M., Zimmermann, M., De Winter, J. C., Bengler, K. J., & Stanton, N. A. (2018). Rolling out the red (and green) carpet: supporting driver decision making in automation-to-manual transitions. *IEEE Transactions on Human-Machine Systems*, *49*(1), 20–31.

Eriksson, A., & Stanton, N. A. (2017). Takeover time in highly automated vehicles: noncritical transitions to and from manual control. *Human factors*, *59*(4), 689–705.

Forster, Y., Hergeth, S., Naujoks, F., Beggiato, M., Krems, J. F., & Keinath, A. (2019). Learning to use automation: Behavioral changes in interaction with automated driving systems. *Transportation research part F: traffic psychology and behaviour*, *62*, 599–614.

Friedman, J. H. (1989). Regularized discriminant analysis. *Journal of the American statistical association*, *84*(405), 165–175.

Gold, C., Happee, R., & Bengler, K. (2018). Modeling take-over performance in level 3 conditionally automated vehicles. *Accident Analysis & Prevention*, *116*, 3–13.

Gold, C., Körber, M., Lechner, D., & Bengler, K. (2016). Taking over control from highly automated vehicles in complex traffic situations: the role of traffic density. *Human factors*, *58*(4), 642–652.

Gold, C., Naujoks, F., Radlmayr, J., Bellem, H., & Jarosch, O. (2017). Testing scenarios for human factors research in level 3 automated vehicles. In *International conference on applied human factors and ergonomics* (pp. 551–559).

ISO. (ISO/TR 21959-1:2020). *Road vehicles — Human performance and state in the context of automated driving — Part 1: Common underlying concepts.*

Jacob, R. J., & Karn, K. S. (2003). Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. In *The mind's eye* (pp. 573–605). Elsevier.

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences*, *105*(19), 6829–6833.

Jones, M., Chapman, P., & Bailey, K. (2014). The influence of image valence on visual attention and perception of risk in drivers. *Accident Analysis & Prevention*, *73*, 296–304.

Keller, J. M., Gray, M. R., & Givens, J. A. (1985). A fuzzy k-nearest neighbor algorithm. *IEEE transactions on systems, man, and cybernetics*(4), 580–585.

Koo, J., Shin, D., Steinert, M., & Leifer, L. (2016). Understanding driver responses to voice alerts of autonomous car operations. *International journal of vehicle design*, *70*(4), 377–392.

Körber, M., Gold, C., Lechner, D., & Bengler, K. (2016). The influence of age on the take-over of vehicle control in highly automated driving. *Transportation research part F: traffic psychology and behaviour*, *39*, 19–32.

Kramer, S. E., Lorens, A., Coninx, F., Zekveld, A. A., Piotrowska, A., & Skarzynski, H. (2013). Processing load during listening: The influence of task characteristics on the pupil response. *Language and cognitive processes*, *28*(4), 426–442.

Lee, J. J., Knox, B., Baumann, J., Breazeal, C., & DeSteno, D. (2013). Computationally modeling interpersonal trust. *Frontiers in psychology*, *4*, 893.

Lee, S.-I., Lee, H., Abbeel, P., & Ng, A. Y. (2006). Efficient l˜ 1 regularized logistic regression. In *Aaai* (Vol. 6, pp. 401–408).

Li, S., Blythe, P., Guo, W., & Namdeo, A. (2018). Investigation of older driver's takeover performance in highly automated vehicles in adverse weather conditions. *IET Intelligent Transport Systems*, *12*(9), 1157–1165.

Luo, R., Wang, Y., Weng, Y., Paul, V., Brudnak, M. J., Jayakumar, P., ... Yang, X. J. (2019). Toward real-time assessment of workload: A bayesian inference approach. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 63, pp. 196–200).

McDonald, A. D., Alambeigi, H., Engström, J., Markkula, G., Vogelpohl, T., Dunne, J., & Yuma, N. (2019). Toward computational simulations of behavior during automated driving takeovers: a review of the empirical and modeling literatures. *Human factors*, *61*(4), 642–688.

McDonald, A. D., Lee, J. D., Schwarz, C., & Brown, T. L. (2014). Steering in a random forest: Ensemble learning for detecting drowsiness-related lane departures. *Human factors*, *56*(5), 986–998.

Mehler, B., Reimer, B., & Coughlin, J. F. (2012). Sensitivity of physiological measures for detecting systematic variations in cognitive demand from a working memory task: an on-road study across three age groups. *Human factors*, *54*(3), 396–412.

Mehler, B., Reimer, B., Coughlin, J. F., & Dusek, J. A. (2009). Impact of incremental increases in cognitive workload on physiological arousal and performance in young adult drivers. *Transportation Research Record*, *2138*(1), 6–12.

Miller, D., Johns, M., Mok, B., Gowda, N., Sirkin, D., Lee, K., & Ju, W. (2016). Behavioral measurement of trust in automation: the trust fall. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 60, pp. 1849–1853).

Mok, B., Johns, M., Lee, K. J., Miller, D., Sirkin, D., Ive, P., & Ju, W. (2015). Emergency, automation off: Unstructured transition timing for distracted drivers of automated vehicles. In *2015 ieee 18th international conference on intelligent transportation systems* (pp. 2458–2464).

Molnar, L. J. (2017). *Age-related differences in driver behavior associated with automated vehicles and the transfer of control between automated and manual* *control: a simulator evaluation* (Tech. Rep.). University of Michigan, Ann Arbor, Transportation Research Institute.

Molnar, L. J., Ryan, L. H., Pradhan, A. K., Eby, D. W., Louis, R. M. S., & Zakrajsek, J. S. (2018). Understanding trust and acceptance of automated vehicles: An exploratory simulator study of transfer of control between automated and manual driving. *Transportation research part F: traffic psychology and behaviour*, *58*, 319–328.

Naujoks, F., Mai, C., & Neukum, A. (2014). The effect of urgency of take-over requests during highly automated driving under distraction conditions. *Advances in Human Aspects of Transportation*, *7*(Part I), 431.

Petersen, L., Robert, L., Yang, J., & Tilbury, D. (2019). Situational awareness, driver's trust in automated driving systems and secondary task performance. *SAE International Journal of Connected and Autonomous Vehicles, 2(2), DOI:10.4271/12-02-02-0009*.

Prusa, J., Khoshgoftaar, T. M., Dittman, D. J., & Napolitano, A. (2015). Using random undersampling to alleviate class imbalance on tweet sentiment data. In *2015 ieee international conference on information reuse and integration* (pp. 197–202).

Quinlan, J. R., et al. (1996). Bagging, boosting, and c4. 5. In *Aaai/iaai, vol. 1* (pp. 725–730).

Radlmayr, J., Gold, C., Lorenz, L., Farid, M., & Bengler, K. (2014). How traffic situations and non-driving related tasks affect the take-over quality in highly automated driving. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 58, pp. 2063–2067).

Ratwani, R. M., McCurry, J. M., & Trafton, J. G. (2010). Single operator, multiple robots: an eye movement based theoretic model of operator situation awareness. In *2010 5th acm/ieee international conference on human-robot interaction (hri)* (pp. 235–242).

Rezvani, T., Driggs-Campbell, K., Sadigh, D., Sastry, S. S., Seshia, S. A., & Bajcsy, R. (2016). Towards trustworthy automation: User interfaces that convey internal and external awareness. In *2016 ieee 19th international conference on intelligent transportation systems (itsc)* (pp. 682–688).

Rish, I., et al. (2001). An empirical study of the naive bayes classifier. In *Ijcai 2001 workshop on empirical methods in artificial intelligence* (Vol. 3, pp. 41–46).

Schmidt, E., Decke, R., & Rasshofer, R. (2016). Correlation between subjective driver state measures and psychophysiological and vehicular data in simulated driving. In *2016 ieee intelligent vehicles symposium (iv)* (pp. 1380–1385).

Society of Automotive Engineers. (2018). *Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems.*

Solovey, E. T., Zec, M., Garcia Perez, E. A., Reimer, B., & Mehler, B. (2014). Classifying driver workload using physiological and driving performance data: two field studies. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 4057–4066).

van den Beukel, A. P., & van der Voort, M. C. (2013). The influence of time-criticality on situation awareness when retrieving human control after automated driving. In *16th international ieee conference on intelligent transportation systems (itsc 2013)* (pp. 2000–2005).

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC bioinformatics*, *7*(1), 91.

Wan, J., & Wu, C. (2018). The effects of lead time of take-over request and nondriving tasks on taking-over control of automated vehicles. *IEEE Transactions on Human-Machine Systems*(99), 1–10.

Wandtner, B., Schömig, N., & Schmidt, G. (2018). Effects of non-driving related task modalities on takeover performance in highly automated driving. *Human factors*, *60*(6), 870–881.

Wang, Y., Reimer, B., Dobres, J., & Mehler, B. (2014). The sensitivity of different methodologies for characterizing drivers' gaze concentration under increased cognitive demand. *Transportation research part F: traffic psychology and behaviour*, *26*, 227–237.

Wintersberger, P., Riener, A., Schartmüller, C., Frison, A.-K., & Weigl, K. (2018). Let me finish before i take over: Towards attention aware device integration in highly automated vehicles. In *Proceedings of the 10th international conference on automotive user interfaces and interactive vehicular applications* (pp. 53–65).

Young, K. L., Salmon, P. M., & Cornelissen, M. (2013). Missing links? the effects of distraction on driver situation awareness. *Safety science*, *56*, 36–43.

Zeeb, K., Buchner, A., & Schrauf, M. (2016). Is take-over time all that matters? the impact of visual-cognitive load on driver take-over quality after conditionally automated driving. *Accident analysis & prevention*, *92*, 230–239.

Zeeb, K., Härtel, M., Buchner, A., & Schrauf, M. (2017). Why is steering not the same as braking? the impact of non-driving related tasks on lateral and longitudinal driver interventions during conditionally automated driving. *Transportation research part F: traffic psychology and behaviour*, *50*, 65–79.

Zhou, F., Alsaid, A., Blommer, M., Curry, R., Swaminathan, R., Kochhar, D., … Lei, B. (2020). Driver fatigue transition prediction in highly automated driving using physiological features. *Expert Systems with Applications*, 113204.

Zhou, F., Yang, X. J., & Zhang, X. (2020). Takeover transition in autonomous vehicles: a youtube study. *International Journal of Human–Computer Interaction*, *36*(3), 295–306.