

# Predicting Driver Errors during Automated Vehicle Takeovers

Hananeh Alambeigi<sup>1</sup> , Anthony D. McDonald<sup>1</sup> , Michael Manser<sup>2</sup>,  
Eva Shipp<sup>2</sup> , John Lenneman<sup>3</sup> , Elizabeth M. Pulver<sup>4</sup>,  
and Scott Christensen<sup>5</sup> 

Transportation Research Record  
1–11

© National Academy of Sciences:  
Transportation Research Board 2023  
Article reuse guidelines:

sagepub.com/journals-permissions  
DOI: 10.1177/03611981231159122

journals.sagepub.com/home/trr



## Abstract

The transition of control between partially automated vehicles and drivers is an important part of the operational design domain and poses unique and important design issues. One approach for enhancing the design of the transition of control mechanisms is to predict driver behavior during a takeover by analyzing his/her state before a takeover. Although there is a wealth of existing literature on modeling the prediction of driver behavior, little is known about the prediction of takeover performance (e.g., driver error) and its underlying data structure (e.g., window sizes or the inclusion of certain features). Thus, the goal of this study is to predict driver error during a takeover event using supervised machine learning algorithms for various window sizes. Three machine learning algorithms (i.e., decision tree, random forest, and support vector machine with a radial basis kernel) were applied to granular driving performance, physiological, and glance data from a driving simulator experiment examining automated vehicle driving. The results showed that a random forest algorithm with an area under the receiver operating characteristic curve of 0.72, trained on a 3 s window before the takeover time, had the best performance with regard to classifying driver error accurately. In addition, we identified the 10 most important predictors that resulted in the best error prediction performance. The results of this study could be useful in developing algorithms for driver state that could be integrated into highly automated systems and, potentially, improve the takeover process.

## Keywords

driver behavior, automated driving, transfer of control, machine learning algorithm, predictive modeling, physiological measures

Automated vehicle technologies have the potential to reduce the nearly six million motor vehicle crashes per year in the U.S.A. (1). However, this potential is limited by the complexity of the interactions between the driver and the automated system (2, 3). In particular, major challenges with regard to safety may arise when drivers are required to take over control of the vehicle when the automated system fails or encounters an operational limit. There is a wealth of existing research that has investigated the factors that influence takeover performance (4), and it suggests that giving drivers more time to react to takeovers and providing them with assistive technology to aid their decision-making during takeovers may improve takeover performance and, consequently, increase safety (4). Machine learning algorithms that accurately predict takeover behavior are an important first step in developing such technology. Although a substantial amount of research has been conducted on

machine learning in the automated driving domain (4), literature in the area of driver takeover during automated driving is still relatively sparse.

Previous studies have predicted driver takeover performance (5–8), situation awareness (9), and fatigue (10) using various machine learning algorithms. Du et al. (5) used six machine learning methods to predict driver takeover performance and categorize it as bad or good according to the takeover reaction time, maximum

<sup>1</sup>Wm Michael Barnes '64 Department of Industrial and Systems Engineering, Texas A&M University, College Station, TX

<sup>2</sup>Texas A&M Transportation Institute, College Station, TX

<sup>3</sup>Toyota Motor North America, Ann Arbor, MI

<sup>4</sup>State Farm Mutual Automobile Insurance Company, Bloomington, IL

<sup>5</sup>State Farm Mutual Automobile Insurance Company, Salem, OR

## Corresponding Author:

Hananeh Alambeigi, hana.alambeigi@tamu.edu

**Table 1.** Ground Truth, Measures, and Algorithms from the Literature and the Current Study

Study	Ground truth	Measures	Algorithm
Du et al. (5)	Takeover performance (good/bad)	Environmental parameters Physiology Eye glance	SVM NB DA kNN LR RF
Ayoub et al. (6)	Takeover performance (time)	Environmental parameters Demographics	XGBoost
Braunagel et al. (7)	Takeover readiness (low/high)	Environmental parameters Eye glance	Linear SVM SVM with a radial basis kernel NB DA kNN
Tivesten et al. (8)	Takeover performance (crash/no crash)	Environmental parameters Eye glance	Metric- and threshold-based
Zhou et al. (9)	Situation awareness (0 to 1)	Eye glance	LightGBM
Zhou et al. (10)	Fatigue (fatigue/nonfatigue)	Physiology Eye glance	NARX RF
Current	Takeover performance (takeover error/no error)	Physiology Eye glance Driving	SVM RBF RF DT

Note: SVM = support vector machine; NB = naïve Bayes; DA = discriminant analysis; kNN = k-nearest neighbors; LR = logistic regression; RF = random forest; XGBoost = eXtreme Gradient Boosting; LightGBM = light gradient boosting method; NARX = nonlinear autoregressive eXogenous network; RBF = radial basis function; DT = decision tree

resulting acceleration, minimum time to collision (TTC), and standard deviation of road offset. They found that the random forest algorithm on a 3 s time window (before the event onset) performed the best when drivers were engaged in nondriving-related tasks. In another study, Ayoub et al. (6) employed eXtreme Gradient Boosting to predict the takeover time using variables that influenced this, for example, the level of automation and the takeover request modality. The analysis found that the urgency of the situation (low, medium, high), takeover time budget, driver's age, and type of nondriving-related task (handheld versus nonhandheld) were the most important variables for predicting takeover time. In a study by Braunagel et al. (7), takeover readiness—an indicator of takeover quality—was categorized as low or high and predicted by three categories of features: complexity of the traffic situation; type of secondary task performed by the driver; and on-road gaze. The study compared support vector machine (SVM) with a linear and radial basis kernel, linear discriminant, naïve Bayes, and k-nearest neighbor (kNN) and found that the SVM with a linear kernel had the highest classification performance. Tivesten et al. (8) developed a simple metric- and threshold-based classifier (i.e., a manual approach for selecting metrics and thresholds that can capture the crash involvement) to predict driver takeover performance categorized as crash and noncrash. This study analyzed driver glance behavior (e.g., number of on-road

and off-road glances) and environmental parameters (e.g., number of warnings issued) and found that a low level of visual attention to the forward roadway, the percentage of time the driver looks on the road during the complete drive, and long visual reaction time to attention reminders are associated with increased risk of crash involvement. Zhou et al. (9) used the light gradient boosting method to predict situation awareness—defined between 0 and 1 based on three performance measures of situation awareness when simulating driving scenarios during the takeover period. This study used eye tracking (e.g., number of fixations on the mirrors) and subjective data (e.g., years of driving experience) as input, and found that features such as the length of the video, the time needed to make a decision, and rearview mirror fixation were the most important in predicting situation awareness. Zhou et al. (10) predicted the driver's transition from nonfatigue to fatigue when driving in automated mode using a random forest algorithm and driver physiology. This analysis found that heart rate, heart rate variability, breathing rate, and standard deviation of breathing rate were the most important features in predicting fatigue. The ground truth, measures, and algorithms used in these studies as well as the current study are summarized in Table 1.

Although these studies provide valuable insights into driver behavior predictions during automated driving, they do not include granular driving performance



**Figure 1.** The driving simulation lab setup. The left figure shows the driver's seat and forward view screens, and the right shows the dashboard and the automated system console (the tablet screen to the right of the steering wheel). Note that the eye-tracking system is positioned on top of the dashboard.

measures such as speed or acceleration, which may be important indicators of successful takeover performance. Furthermore, the ground truth definitions of takeover errors in previous work are grounded in driving performance rather than execution of the takeover action. There is also a need to replicate the findings of these studies to understand algorithm characteristics and performance relationships that can be replicated across datasets. The goal of this study is to expand the previous analyses to predict driver error during a takeover process using machine learning algorithms for a range of window sizes using a set of data generated from a driving simulator experiment during automated vehicle driving. In particular, this study aims to understand the importance of including granular driving performance data in combination with physiological and glance data for predicting takeover errors.

## Methods

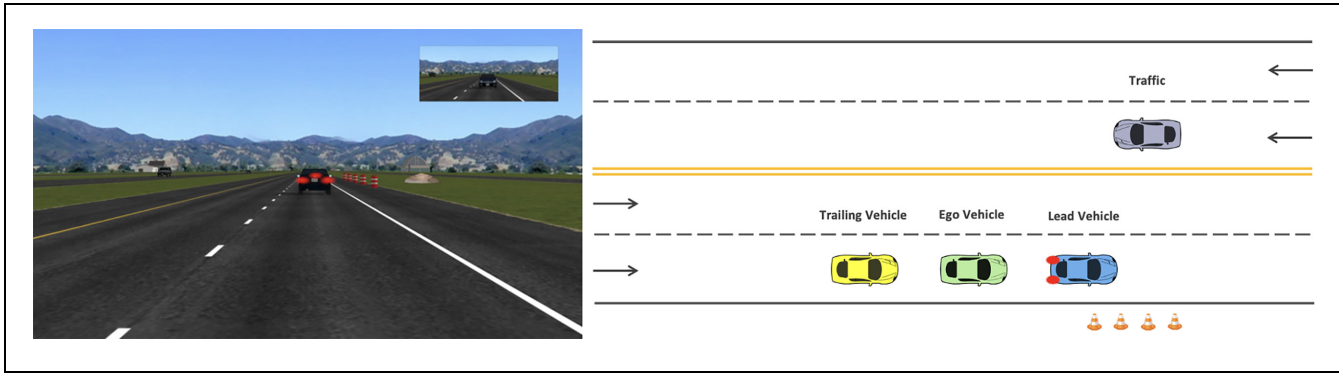
The driving simulation experiment data were collected in a Realtime Technologies driving simulator lab at the Texas A&M Transportation Institute. The lab consists of a quarter-cab driving simulator with three screens that provide 165° horizontal and 35° vertical fields of view, a speaker system to provide ambient roadway noise, and a physiological and eye-tracking data collection suite. The driving simulator setup is illustrated in Figure 1. The original goal of the study was to analyze driver behavior after silent automation failures and develop models of driver behavior. The driving environment and automated driving system were simulated using SimCreator software, and were in line with level 2 automation according to the Society of Automotive Engineers (SAE);

specifically, the software was configured with adaptive cruise control and lane-keeping ability. The simulator's automated driving system was activated with a button on a touch screen display located to the right of the steering wheel. When the system encountered a failure or an operational limit, the vehicle's automated system was disabled (see Alambeigi and McDonald (11) for a detailed description).

## Dataset

The study involved 64 participants (32 males, 32 females) aged 19 to 65 with a mean age of 41.44 (SD = 15.14) years from the surrounding community. All participants were English speakers, reported normal or corrected-to-normal visual acuity and normal color vision, held a valid driver's license, reported driving experience of at least 1.5 years, were not on any medication that may have affected their ability to operate a moving vehicle, had not previously participated in an experiment involving automated vehicles, and had no previous experience of driving an automated vehicle (SAE level 2 or higher). All procedures were approved by the Texas A&M Institutional Review Board (IRB2018-1362D) and were conducted in accordance with the principles expressed in the American Psychological Association Code of Ethics. Informed consent was obtained from each participant and they received \$50 for taking part.

Throughout the experiment, driving performance data, including continuous steering wheel position, accelerator and brake pedal positions, velocity, time to lane crossing, time headway to an upstream object, and lane position were collected at a 60 Hz sampling rate. Physiological indicators, including heart rate, breathing



**Figure 2.** Unexpected braking takeover scenario with the construction zone on the road shoulder. The left figure shows the simulator scenario from the driver's view and the right figure shows the scenario schematic from the top view.

rate, and electrodermal activity (EDA), were also collected from each participant. Heart rate and breathing rate were measured using a Zephyr BioHarness 3.0 (Zephyr Technology, Annapolis, MD), which is an adjustable chest strap with an embedded ECG sensor, at a 1 Hz sampling rate. The EDA data were measured at 60 Hz using a Shimmer3 wireless galvanic skin response (GSR) sensor (Shimmer, Dublin, Ireland), which is an elastic strap that was attached to subjects' wrists on their nondominant hand, and two electrodes attached to the palm. Glance behavior data were collected using a dashboard-mounted FOVIO eye-tracking system (Seeing Machines, Canberra, Australia), which was interfaced with EyeWorks data recordings software (EyeTracking LLC, Solana Beach, CA). Participants were calibrated to the FOVIO system using a four-point calibration screen and were instructed to look at the exterior edges of the panoramic display while maintaining a directly forward field of view. The driving performance dataset compiled from the experiment has been published in the Virginia Tech Transportation Institute data repository (12).

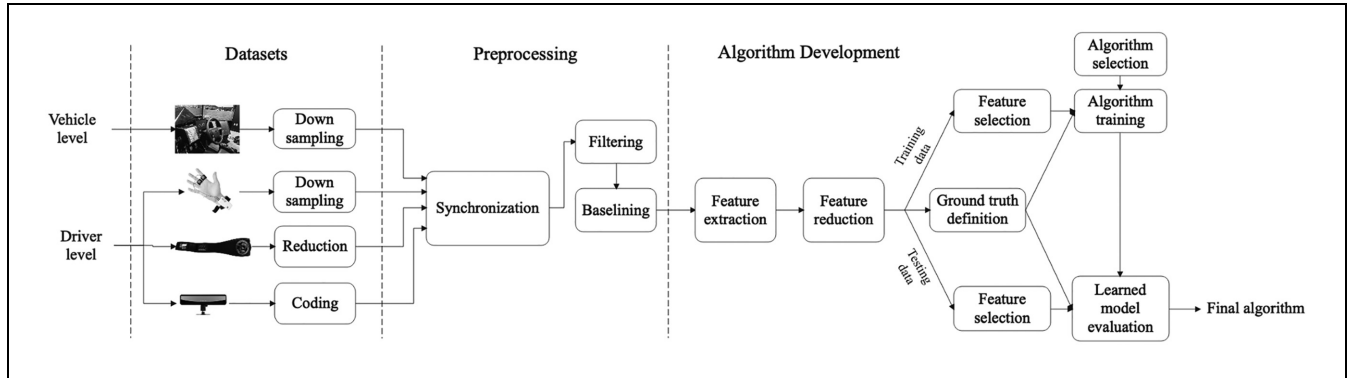
The study process consisted of a 10-min training session on the system capability and operation, two practice drives, and four counterbalanced experimental drives separated by 2-min breaks. The experimental portion of the study was designed as a  $2 \times 2 \times 2$  study with scenario criticality (critical or noncritical) and failure type (braking response or unexpected obstacle) as within-subjects factors and alert type (silent, audiovisual) as a between-subjects factor. For the purposes of the current analysis, only the braking response failure is included, because it has been found to be one of the most commonly experienced real-world scenarios (3). The experimental drives took place on a 10-mile section of a 4-lane straight highway on which the participants drove in a 3-vehicle platoon with a 1 s time headway. The unexpected braking scenario included two braking events after approximately three and seven miles of driving. In the first event, the automated system responded

appropriately when the lead vehicle braked and in the second event the automated system failed to respond. In the latter event, the vehicle's lateral and longitudinal control failed, necessitating a takeover. The criticality of the scenario was manipulated using the deceleration rates of the lead vehicle, for which the constant deceleration rate of  $2 \text{ m/s}^2$  represented the noncritical scenario and  $5 \text{ m/s}^2$  represented the critical scenario. Figure 2 shows the unexpected braking takeover scenario from the driver's view (left) and an overhead view (right). At the start of the failure in each drive, participants in the audiovisual alert type condition received an auditory alert (loud beeping sound) and a visual alert (change of color on the instrument cluster and a notification displayed on the automated system activation screen), indicating the need to take over. Participants in the silent failure condition received no alert. Drivers were instructed to keep their hands on the steering wheel throughout the experiment and informed that it was their responsibility to monitor the automated system and the driving environment.

### Data Preprocessing and Ground Truth Definition

All 64 participants completed the entire experiment, resulting in 128 completed driving performance, physiological, and glance datasets; however, physiological data from the BioHarness—including the heart rate variability—for four drivers were missing and, thus, excluded. Two additional participants were excluded from the datasets because of technical and calibration issues in relation to the eye tracking, resulting in 122 complete datasets. All the data preprocessing and analysis steps were performed in R 4.0.3 (13) using the “tidy-models” package (14). Figure 3 illustrates the entire analysis schematic.

The driving and physiological datasets started from the beginning of the drive and ended approximately 3 min after the event was completed. The driving performance data and the physiological data from the GSR



**Figure 3.** Analysis schematic, including the datasets, preprocessing steps, and development of the algorithm

sensor were down-sampled to 10 Hz from 60 Hz using the mean of every six samples. The BioHarness yielded 1 Hz data from the entire experiment process and these were assigned to each scenario based on the synchronized time. The glance data were manually annotated by two independent coders from 10 s before the event onset until the end of the event. The event onset was defined as the time when the lead vehicle started decelerating and the automated system failed to respond, and the end of the event was defined as the time at which either a crash happened or the situation was resolved (i.e., the time at which the driver released the brake then accelerated, or completed the lane change maneuver by stabilizing the vehicle in the new lane). The areas of interest in the coding process included glance at the lead vehicle, dashboard, automated system console, construction site, road, and off road (e.g., surrounding buildings). For the purposes of the current analysis, the driving performance data included only the time range during which the vehicle was manually driven by the driver following the failure onset until the time of the takeover, the physiological data included the entire drive up to the time of the takeover, and the glance data included 10 s before the event to the time of the takeover. The takeover time was considered as the time between the event onset and the start of a braking maneuver, steering maneuver, or both, greater than a certain threshold. The data from all these sources were time synced to a 10<sup>th</sup> of a millisecond.

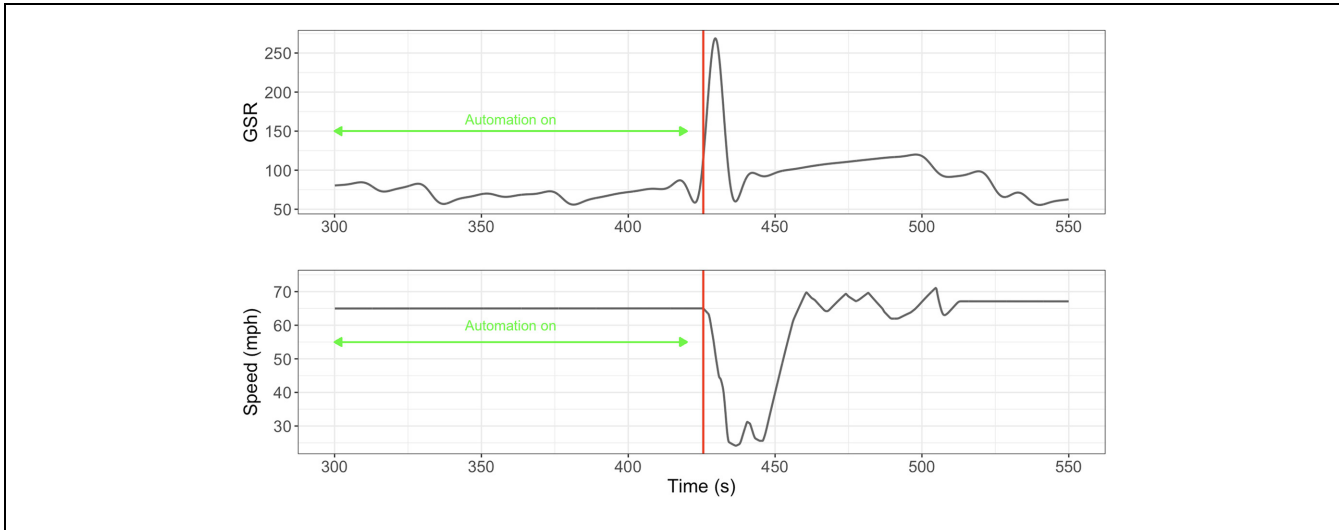
After the data were integrated, a data filtering and baselining process was performed. First, a plausibility filter was applied to the physiological data to remove invalid data (e.g., heart rate values of 0) that were a result of the posture adopted by the participants, which made the chest strap sensor slide against the participant's skin and lose contact, a result of poor fitting. This step was guided by the data recording limits in each device's user manual. Next, a low-pass Butterworth filter with a sampling frequency of 1 Hz and a cutoff frequency of 0.1 Hz was applied to reduce noise. The optimal cutoff frequency was computed following the work by Yu et al. (15). Once the

noise removal process was completed, the physiological data were scaled relative to a baseline, which was defined as the mean of a 30-s time period of automated driving from the beginning of the drive after enabling the automated driving system and before encountering the event for each participant. The selection of this method for defining the baseline was guided by previous driving simulator studies (16, 17). An example of the processed physiological and driving performance data is shown in Figure 4.

The data were labeled as either error or no error based on driver performance during the takeover process. Error was defined as a failure to complete a necessary subtask during the takeover maneuver (e.g., failing to check the side mirror before a lane change) or completing the necessary tasks but in the wrong order (e.g., checking the side mirror after a lane change). Eventually, 22 drives were labeled as error and 100 drives as no error. Table 2 shows the order of subtasks associated with a braking or a lane change maneuver, and the categories used to define an error.

### Feature Extraction and Reduction

Following data preprocessing, a set of 73 features was extracted for window sizes including 3 s, 5 s, 10 s, 15 s, 20 s, 30 s, 60 s, 120 s, and 300 s before the takeover time. The driving features were limited to post-event and glance features were limited to 10 s before the event. Thus, longer window sizes (>10 s) mostly consisted of physiological features. The takeover time was defined as the time between the event onset and the start of the maneuver greater than a threshold of 2° for steering wheel angle rotation and 10% for brake pedal position (4). Features were generated from the driving performance, physiological, and glance measures. Table 3 shows the features extracted along with their corresponding measures and data sources. After the features were generated, they were centered and scaled and feature reduction was performed to remove features with near-zero variance (cutoff value of 19) and highly correlated data (features with an absolute Pearson correlation



**Figure 4.** An example of the preprocessed data. The top plot shows the GSR from the physiological dataset and the bottom plot shows the speed of the vehicle from the driving simulator dataset.

Note: GSR = galvanic skin response

**Table 2.** Order of Subtasks Associated with a Braking or Lane-Changing Maneuver and the Categories Used to Define an Error

Maneuver	Subtask	Error
Braking	Looking at the lead vehicle	Not checking the rearview mirror
	Moving hands/feet toward the wheel/pedal	Braking before checking the rearview mirror
	Checking the rearview mirror	Crash
	Applying the brake	
	Avoiding a crash	
Lane changing	Looking at the lead vehicle	Not checking the sideview mirror
	Moving hands/feet toward the wheel/pedal	Lane changing before checking the sideview mirror
	Checking the sideview mirror	Driving off the road
	Applying the brake	Crash
	Avoiding a crash	

greater than 0.85). Lastly, the data were up-sampled. These steps were guided by the work in McDonald et al. (18). The feature reduction resulted in a total of 42 features in each window size. These features are highlighted in Table 3.

### Algorithm Training and Evaluation

Three machine learning algorithms—decision tree, random forest, and SVM with a radial basis kernel—were trained for each of the window size datasets. Each dataset contained one row for each drive and 42 features defined for the given window size. These algorithms were selected because they are the most commonly applied methods in the field and they allowed a comparison between simple and complex models. Although other approaches also have several advantages, the study design limited the use of these algorithms (e.g., insufficient data to use neural

network and deep neural network). In addition, previous research suggests that logistic regression kNN was unlikely to outperform SVM and tree-based models (5). Down-selecting the algorithms also helps to maintain statistical power and avoid an excessive number of pairwise comparisons in performing statistical analysis. The training process consisted of a five-fold grouped cross-validation process. The data were partitioned at the driver level (to avoid a driver's dataset being included in both training and testing). Following data partitioning, the data were up-sampled to create a balanced training set. The trained algorithms were assessed by their area under the receiver operating characteristic (ROC) curve (AUC) across the five groups (19), for which a higher value of AUC indicates a better performance. The algorithms' AUC differences were statistically evaluated using the DeLong test for ROC curves with a threshold of  $p < .05$ . The DeLong test is a nonparametric test that

**Table 3.** Categorization of the Datasets, Measures, and the Extracted Features

Data source	Measure	Unit	Feature
Driving simulator	Longitudinal and lateral speed	Meters per second	<b>Max</b> , min, mean, <b>med</b> , and SD of the speed
	Longitudinal and lateral acceleration	Meters per squared second	<b>Max</b> , min, mean, med, and SD of the acceleration
	Acceleration and brake pedal position	na	<b>Max</b> , min, mean, <b>med</b> , <b>SD</b> , and <b>zero crossing rate</b> of the pedal position
	Lane offset	Inches	<b>Max</b> , <b>min</b> , mean, <b>med</b> , SD, and <b>lane center crossing rate</b>
	Steering wheel angle	Degrees	<b>Max</b> , <b>min</b> , mean, <b>med</b> , <b>SD</b> , <b>zero crossing rate</b> , maximum steering wheel angle rate, and <b>sample entropy</b> of the steering wheel angle
	Automation disengagement	Count	<b>Rate</b> of disengagement
	Time to collision	Seconds	<b>Min</b> TTC after the event onset
BioHarness/GSR	Heart rate	Beats per min	<b>Max</b> , <b>min</b> , mean, med, and <b>SD</b> of heart rate
	Heart rate variability	Standard deviation in milliseconds	<b>Max</b> , min, mean, med, and <b>SD</b> of heart rate variability
	Breathing rate	Breaths per minute	<b>Max</b> , <b>min</b> , mean, med, and <b>SD</b> of breathing rate
	Galvanic skin response	Kilo ohms	<b>Max</b> , min, mean, med, and <b>SD</b> of electrodermal activity
FOVIO	First fixation location	na	<b>Location of the first observed area of interest</b> after the event onset
	First fixation duration	Seconds	<b>Duration of the first observed area of interest</b> after the event onset
	Fixation rate	Count	<b>Number of fixations on areas of interests</b>
	Fixation change rate	Count	<b>Number of changes in fixation location</b>
	Eyes off road	Seconds	Duration of off-road glances
	Eyes on road	Seconds	<b>Duration of on-road glances</b>

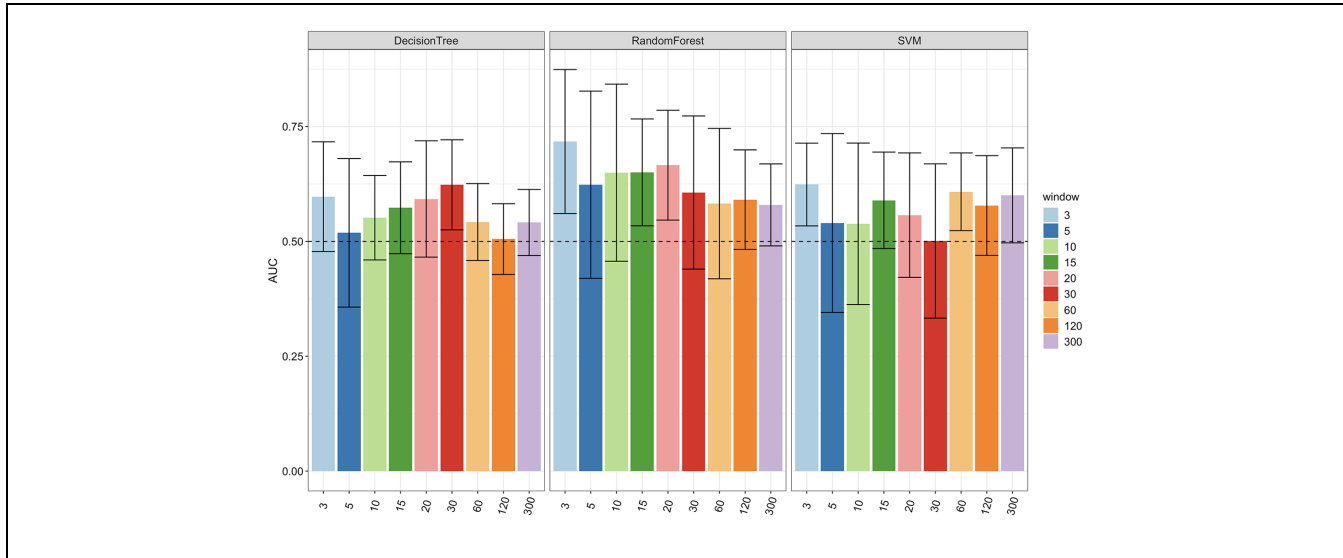
Note: max = maximum; min = minimum; med = median; SD = standard deviation; TTC = time to collision; GSR = galvanic skin response; na = not applicable. The features included in the algorithms are highlighted in bold.

can be used to investigate whether the AUCs of two models are statistically significantly different. The DeLong test on the ROC curve has been shown to be equal to the Mann–Whitney U-statistic for comparing distributions of values from the two samples. In addition to the analysis of algorithm performance, a feature importance analysis using permutation-based importance measure was performed for the algorithm with the best predictive performance to provide additional insights into the drivers' behavioral patterns. The feature importance values indicate each feature's predictive ability by computing mean decrease in accuracy in predicting an error. In this analysis, larger values of decrease in accuracy represent the most important features. Therefore, removing a variable from the model lessens the accuracy of that particular model.

## Results

Figure 5 shows the algorithm AUC categorized by the window size and machine learning algorithm (decision tree, random forest, SVM with a radial basis kernel). The black error bars indicate the 95% confidence intervals based on

the standard errors. The dashed line in Figure 5 shows random guessing performance. The statistical analysis of the algorithms showed that the random forest algorithm outperformed the other models across most of the window sizes. Therefore, random forest was selected for further analysis. This is in line with previous research that found random forest generally outperforms simple decision tree and boosted tree models. Figure 5 shows that the random forest algorithm with a 3 s window size had the highest AUC of 0.72 with the 95% confidence interval of (0.56, 0.87) followed by a 20 s window with an AUC of 0.67 (0.55, 0.78) and then a 15 s window with an AUC of 0.65 (0.55, 0.77) both from the random forest. A significant difference was found in the AUC between the random forest with a 3 s time window and random guessing ( $p = .01$ ). In addition, pairwise comparison showed that random guessing outperformed random forest for 5 s, 30 s, 60 s, 120 s, and 300 s window sizes ( $p < .05$ ). However, no significant differences were found between 3 s and 10 s ( $p = .40$ ), 15 s ( $p = .26$ ), and 20 s ( $p = .50$ ) windows for random forest. Table 4 shows the results of the DeLong test for the pairwise comparisons between the random forest algorithm with a 3 s window size and the other fitted algorithms.



**Figure 5.** Algorithm AUC categorized by machine learning approach and window size. The horizontal dashed line indicates random guessing performance. The error bars indicate the 95% confidence intervals based on the standard errors.  
 Note: AUC = area under the curve; SVM = support vector machine

**Table 4.** DeLong Test Results for Pairwise Comparisons between the Random Forest Algorithm with a 3 s Window Size and the Other Fitted Algorithms.

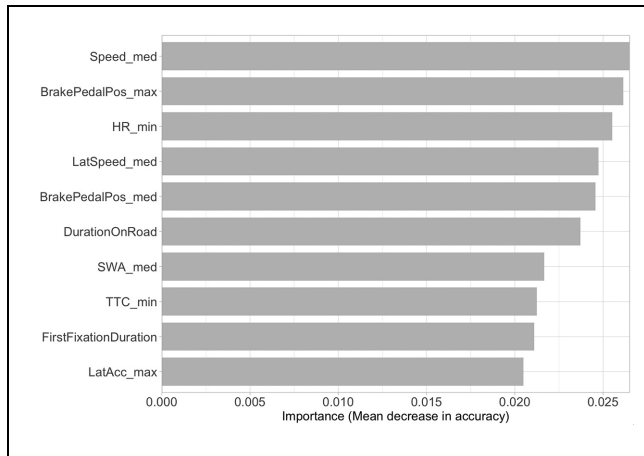
Classification algorithm	Window size	AUC (bootstrapped confidence interval)	DeLong test (random forest 3 s)
Random forest	3	0.72 (0.56, 0.87)	na
	5	0.62 (0.42, 0.82)	D(225.50)=3.30; $p < 0.05$
	10	0.65 (0.46, 0.83)	D(225.90)=0.67; $p = 0.40$
	15	0.65 (0.55, 0.77)	D(226.22)=0.68; $p = 0.26$
	20	0.67 (0.55, 0.78)	D(218.60)=0.49; $p = 0.50$
	30	0.61 (0.45, 0.77)	D(221.10)=3.08; $p < 0.05$
	60	0.59 (0.44, 0.74)	D(226.49)=3.04; $p < 0.05$
	120	0.59 (0.47, 0.71)	D(225.91)=3.06; $p < 0.05$
	300	0.58 (0.49, 0.67)	D(224.29)=2.51; $p < 0.05$
Decision tree	3	0.59 (0.48, 0.72)	D(225.59)=2.66; $p < 0.05$
	5	0.51 (0.36, 0.67)	D(223.23)=2.76; $p < 0.05$
	10	0.55 (0.46, 0.64)	D(223.50)=2.87; $p < 0.05$
	15	0.57 (0.47, 0.67)	D(219.37)=3.00; $p < 0.05$
	20	0.58 (0.46, 0.71)	D(219.00)=3.01; $p < 0.05$
	30	0.62 (0.53, 0.71)	D(221.11)=2.14; $p < 0.05$
	60	0.54 (0.46, 0.62)	D(224.41)=2.58; $p < 0.05$
	120	0.50 (0.42, 0.58)	D(222.84)=2.54; $p < 0.05$
	300	0.54 (0.47, 0.61)	D(225.59)=2.66; $p < 0.05$
SVM	3	0.62 (0.53, 0.70)	D(225.15)=2.60; $p < 0.05$
	5	0.53 (0.33, 0.72)	D(218.35)=2.64; $p < 0.05$
	10	0.52 (0.35, 0.70)	D(210.48)=2.85; $p < 0.05$
	15	0.58 (0.48, 0.69)	D(211.77)=3.00; $p < 0.05$
	20	0.55 (0.42, 0.68)	D(221.89)=2.60; $p < 0.05$
	30	0.50 (0.35, 0.64)	D(221.86)=2.99; $p < 0.05$
	60	0.60 (0.52, 0.68)	D(206.46)=3.54; $p < 0.05$
	120	0.56 (0.45, 0.67)	D(226.00)=2.14; $p < 0.05$
	300	0.60 (0.49, 0.70)	D(217.67)=3.39; $p < 0.05$

Note: AUC = area under the curve; SVM = support vector machine; na = not applicable

In addition to the algorithm analysis, feature importance values were computed to provide additional insight into each feature’s relative importance in the takeover

error prediction. The importance values indicate each feature’s mean decrease in accuracy in predicting the error. Thus, larger values of decrease in accuracy





**Figure 6.** Mean decrease in algorithm accuracy of the 3 s window random forest algorithm associated with the top 10 features

Note: med = median; max = maximum; min = minimum; Pos = position; HR = heart rate; Lat = lateral; SWA = steering wheel angle; TTC = time to collision; Acc = acceleration

represent the most important features. Figure 6 shows the 10 most important features for the random forest model with a 3 s time window. The results show the importance of all three measures used: granular driving performance, physiological, and glance data. The figure indicates that median speed is the most important feature, although features derived from heart rate, glance duration, braking, and steering behavior are also important. This is worthy of note because of the potential implications for data collection requirements in relation to future error prediction technology.

## Discussion

In this study, we developed 27 machine learning algorithms to predict takeover errors using a set of granular driving performance, physiological, and glance data that were gathered during partially automated vehicle driving. The 27 algorithms differed according to the machine learning approach (random forest, decision tree, SVM with a radial basis kernel) and the window size of data before takeover (3 s, 5 s, 10 s, 15 s, 20 s, 30 s, 60 s, 120 s, 300 s). The results suggest that these algorithms can predict takeover errors significantly better than random guessing, although the findings are inconsistent. There is some indication that a window size of 3 s leads to a higher AUC for the random forest algorithm; however, the difference was not significant across some other window sizes (i.e., 10 s, 15 s, and 20 s).

The results showed that the random forest classifier outperformed the remaining algorithms as indicated by the AUC values. This finding is consistent with previous studies in the automated vehicle driving domain. Du

et al. (5) found random forest as a classifier to have the highest mean prediction accuracy (83%) compared with other approaches, including decision tree and SVM. The results of the AUC show a high value for the random forest model with a 3 s time window (0.72), followed by the 20 s (0.67) and 15 s (0.65) windows. A significant difference was shown for the 3 s window and random guessing ( $p = 0.01$ ). In addition, with regard to the random forest model, the size of the window significantly influenced the prediction performance. Pairwise comparisons between 3 s and other window sizes showed significant differences between the 3 s window and 5 s, 30 s, 60 s, 120 s, and 300 s window sizes, whereas no significant differences were found for the 10 s, 15 s, and 20 s sizes. Despite the fact there were no significant differences between the random forest with 3 s and the random forest with 10 s, 15 s, and 20 s window sizes, we focused the remainder of the analysis on the random forest with 3 s for brevity. Notably, there were no differences in the important features across these window sizes. In addition, a 3 s random forest may be preferable in an applied setting because of the reduction in data retention needed for a 3 s feature window compared with 10–20 s windows. This finding aligns with that of Du et al. (5), who recommended 3 s as the optimal timeframe for predicting driver takeover performance. Although a broader range of (physiological) measures—up to 300 s before the takeover—were included in this study, no significant improvement was found. Collectively, these findings might suggest that as we go further from the takeover event (i.e., more than 20 s), the correlations between the takeover error and other influential factors—in particular physiological—fade. Further investigation is needed to explore this speculation.

The analysis of feature importance highlights the necessity of including a combination of driving performance, physiological, and glance measures in takeover error prediction. The findings in relation to the 10 most important features show that driving variables, including median lateral and longitudinal speed, median steering wheel angle, minimum TTC, maximum and median brake pedal position, and maximum lateral acceleration play an important role in error detection. Finding minimum TTC to be one of the most important features is worthy of note. Minimum TTC, which has been used in several studies, is an established surrogate safety metric for longitudinal vehicle control (4). Inverse TTC is associated with the perceived criticality of the situation and has been shown to have a strong link with driver behavior because it may trigger emergency avoidance reactions (20, 21). Thus, a lower minimum TTC might lead to a more abrupt maneuver and, thus, more errors. It is important to note that these measures are more granular than the environmental parameters (e.g., traffic density)

included in previous work (5, 7, 8). Perhaps the most relevant feature in relation to the findings of this study—in particular with regard to the minimum TTC—is the takeover time budget found in Du et al. (5), which was identified as one of the important features in takeover performance prediction. Previous studies have shown that a shorter takeover time budget is associated with a shorter minimum TTC (11). Thus, our finding is aligned with the takeover time budget found in Du et al. (5).

With regard to physiological indices, heart rate was found to be the most important feature. This result aligns with Du et al. (5), who also identified heart rate-based measures as being important. The duration of on-road glances and the duration of first fixation 3 s before the takeover were the most important glance features. This finding might be associated with the visual readiness component of a takeover process in which the driver has to redirect his/her gaze to the forward roadway. Thus, it is reasonable to say that as the duration of on-road glances decreases, the probability of making an error increases. This finding is consistent with Tivesten et al. (8), who found that low visual attention to the forward roadway was associated with an increased risk of being involved in a crash.

The results of this study have implications for developing algorithms for driver error detection and mitigation. The findings suggest that 3 s before the takeover is required to predict driver takeover error, and the driving performance, physiological, and glance measures need to be collected within this range. Given these findings, consideration should be given to designing advanced in-vehicle monitoring systems to monitor the driver's state proactively, and issue a warning if an abnormality is detected. Providing dynamic feedback to the driver can mitigate driver takeover errors.

### Limitations and Future Work

Although the analysis provides useful insights into data implementation and driver error prediction, it is limited in some respects. First, the number of observations in this study was relatively small, which limits the use of more complicated algorithms. In addition, given the nature of the simulator scenario, the driving performance data are only available after the onset of the event to the takeover time, which might lead to inconsistencies between the physiological and glance data, and the length of the window sizes. Moreover, the collected data were from a driving simulator, which provides relative validity and might not reflect real-world situations. Another limitation is that the feature importance analysis only highlighted the most important features and their relative importance for accurate classification; to obtain information with regard to the magnitude of changes in a variable and its impact on prediction,

further analysis such as partial dependence or Shapley values is needed. Future work should focus on analysis of driver error using a larger and more diverse dataset, including varied levels of pre-takeover stress and different driver characteristics, and validate the findings in on-road real-world automated driving settings.

### Conclusions

The goal of the current study was to investigate which driving performance, physiological, and glance measures before a takeover can capture driver error during a takeover process. In addition, the study focused on detecting an effective range of data for implementing the model predictions. We analyzed a combination of physiological, driving performance, and glance data from a driving simulator experiment during a partially automated experimental drive in which the takeover scenario consisted of a lead vehicle unexpectedly braking because it was approaching a construction site. A set of features has been generated, and three machine learning algorithms (SVM with a radial basis kernel, decision tree, and random forest) were applied to these features. We found the random forest with an AUC of 0.72 to be the best classifier for predicting driver error based on a 3 s time window before the takeover time. In addition, the results highlighted the importance of driving performance measures including speed, brake pedal position, TTC, acceleration and steering wheel angle, physiological measures including heart rate, and glance measures including the duration of on-road glances and duration of first fixation. The findings provide useful insights into data collection requirements for designing driver error prediction technologies.

### Author Contributions

The authors confirm contribution to the paper as follows: study conception and design: M. Manser, E. Shipp, A. McDonald; data collection: H. Alambeigi, A. McDonald; analysis and interpretation of results: H. Alambeigi, A. McDonald, M. Manser, E. Shipp, J. Lenneman, E. Pulver, S. Christensen; draft manuscript preparation: H. Alambeigi, A. McDonald, M. Manser, J. Lenneman. All authors reviewed the results and approved the final version of the manuscript.

### Declaration of Conflicting Interests






The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Support for this research was provided in part by grants

from the Toyota Collaborative Safety Research Center, the State Farm Insurance Company, and the U.S. Department of Transportation, University Transportation Centers Program via the Safety through Disruption University Transportation Center (451453-19C36).

### ORCID iDs

Hananeh Alambeigi  <https://orcid.org/0000-0003-4310-3950>  
 Anthony D. McDonald  <https://orcid.org/0000-0001-7827-8828>  
 Eva Shipp  <https://orcid.org/0000-0002-4034-8031>  
 John Lenneman  <https://orcid.org/0000-0002-0502-1690>  
 Scott Christensen  <https://orcid.org/0000-0003-2397-2286>

### References

1. National Highway Traffic Safety Administration. *Police-Reported Motor Vehicle Traffic Crashes in 2018*. Traffic Safety Facts. Research Note. NHTSA, Washington, DC, 2020.
2. Lu, Z., and J. C. F. de Winter. A Review and Framework of Control Authority Transitions in Automated Driving. *Procedia Manufacturing*, Vol. 3, 2015, pp. 2510–2517. <https://doi.org/10.1016/j.promfg.2015.07.513>.
3. Alambeigi, H., A. D. McDonald, and S. R. Tankasala. Crash Themes in Automated Vehicles: A Topic Modeling Analysis of the California Department of Motor Vehicles Automated Vehicle Crash Database. Presented at 99th Annual Meeting of the Transportation Research Board, Washington, DC, 2019. <http://arxiv.org/abs/2001.11087>
4. McDonald, A. D., H. Alambeigi, J. Engström, G. Markkula, T. Vogelpohl, J. Dunne, and N. Yuma. Toward Computational Simulations of Behavior during Automated Driving Takeovers: A Review of the Empirical and Modeling Literatures. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, Vol. 61, No. 4, 2019, pp. 642–688. <https://doi.org/10.1177/0018720819829572>.
5. Du, N., F. Zhou, E. M. Pulver, D. M. Tilbury, L. P. Robert, A. K. Pradhan, and X. J. Yang. Predicting Driver Takeover Performance in Conditionally Automated Driving. *Accident Analysis and Prevention*, Vol. 148, 2020, pp. 1–11. <https://doi.org/10.1016/j.aap.2020.105748>.
6. Ayoub, J., N. Du, X. J. Yang, and F. Zhou. Predicting Driver Takeover Time in Conditionally Automated Driving. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 23, No. 7, 2022, pp. 9580–9589.
7. Braunagel, C., W. Rosenstiel, and E. Kasneci. Ready for Take-Over? A New Driver Assistance System for an Automated Classification of Driver Take-Over Readiness. *IEEE Intelligent Transportation Systems Magazine*, Vol. 9, No. 4, 2017, pp. 10–22. <https://doi.org/10.1109/MITS.2017.2743165>.
8. Tivesten, E., T. W. Victor, P. Gustavsson, J. Johansson, and M. L. Aust. Out-of-the-Loop Crash Prediction: The Automation Expectation Mismatch (AEM) Algorithm. *IET Intelligent Transport Systems*, Vol. 13, No. 8, 2019, pp. 1231–1240. <https://doi.org/10.1049/iet-its.2018.5555>.
9. Zhou, F., X. J. Yang, and J. C. F. de Winter. Using Eye-Tracking Data to Predict Situation Awareness in Real Time during Takeover Transitions in Conditionally Automated Driving. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 23, No. 3, 2021, pp. 2284–2295. <https://doi.org/10.1109/TITS.2021.3069776>.
10. Zhou, F., A. Alsaïd, M. Blommer, R. Curry, R. Swaminathan, D. Kochhar, W. Talamonti, L. Tijerina, and B. Lei. Driver Fatigue Transition Prediction in Highly Automated Driving Using Physiological Features. *Expert Systems with Applications*, Vol. 147, 2020, article 113204. <https://doi.org/10.1016/j.eswa.2020.113204>.
11. Alambeigi, H., and A. D. McDonald. A Bayesian Regression Analysis of the Effects of Alert Presence and Scenario Criticality on Automated Vehicle Takeover Performance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, Vol. 65, No. 2, 2021, pp. 288–305. <https://doi.org/10.1177/00187208211010004>.
12. Alambeigi, H., and T. McDonald. *Investigating the Effects of Silent Automation Failure and Scenario Criticality on Automated Vehicle's Takeover Performance (03-036)*. Virginia Tech Transportation Institute, Blacksburg, VA, 2021.
13. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. <https://www.r-project.org/>.
14. Kuhn, M., and H. Wickham. *Easily Install and Load the "Tidymodels" Packages*, 2022. <https://tidymodels.tidymodels.org>
15. Yu, B., D. Gabriel, L. Noble, and K. N. An. Estimate of the Optimum Cutoff Frequency for the Butterworth Low-Pass Digital Filter. *Journal of Applied Biomechanics*, Vol. 15, No. 3, 1999, pp. 318–329. <https://doi.org/10.1123/jab.15.3.318>.
16. Kim, A. J., H. Alambeigi, T. Goddard, A. D. McDonald, and B. A. Anderson. Bicyclist-evoked arousal and greater attention to bicyclists independently promote safer driving. *Cognitive Research: Principles and Implications*, Vol. 6, Article Number: 66, 2021. <https://doi.org/10.1186/s41235-021-00332-y>
17. Son, J., B. Mehler, T. Lee, Y. Park, J. Coughlin, and B. Reimer. Impact of Cognitive Workload on Physiological Arousal and Performance in Younger and Older Drivers. *Proc., Driving Assessment Conference*, Vol. 6, University of Iowa, June 28, 2011, pp. 87–94. <https://doi.org/10.17077/drivingassessment.1382>.
18. McDonald, A. D., T. K. Ferris, and T. A. Wiener. Classification of Driver Distraction: A Comprehensive Analysis of Feature Generation, Machine Learning, and Input Measures. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, Vol. 62, No. 6, 2020, pp. 1019–1035. <https://doi.org/10.1177/0018720819856454>.
19. Fawcett, T. ROC Graphs: Notes and Practical Considerations for Researchers. *Machine Learning*, Vol. 31, No. 1, 2004, pp. 1–38.
20. Markkula, G., J. Engström, J. Lodin, J. Bårgman, and T. W. Victor. A Farewell to Brake Reaction Times? Kinematics-Dependent Brake Response in Naturalistic Rear-End Emergencies. *Accident Analysis & Prevention*, Vol. 95, 2016, pp. 209–226. <https://doi.org/10.1016/j.aap.2016.07.007>.
21. Engström, J. Scenario Criticality Determines the Effects of Working Memory Load on Brake Response Time. *Proc., European Conference on Human Centred Design for Intelligent Transport Systems*, Lyon, France, April 2010, pp. 25–36.