electronics

*Article*

# Towards Safer Automated Driving: Predicting Drivers with Long Takeover Time Using Random Forest and Human Factors

**Jungsook Kim [1,2] and Ohyun Jo [1,*]**

[1] Department of Computer Science, College of Electrical and Computer Engineering, Chungbuk National University, Cheongju 28644, Republic of Korea; jungsook96@chungbuk.ac.kr
[2] Information Strategy Department, Electronics and Telecommunications Research Institute, Daejeon 34129, Republic of Korea
* Correspondence: ohyunjo@chungbuk.ac.kr

**Abstract**

In highly automated driving systems (ADSs), drivers' ability to resume manual driving remains a road safety issue. However, to the best of our knowledge, there is no existing computational model to predict which drivers require more than the 4 seconds mandated by United Nations Regulation No. 157 to regain manual control. To address this challenge, we developed a Random Forest model that predicts takeover time using measurable human factors. Three controlled driving simulator experiments were conducted in which participants engaged in distinct tasks—texting, drinking, and traffic monitoring—before responding to a takeover request. During the experiments, we collected human factor features, including gaze behavior, age, and scores, from the self-reported driving behavior questionnaire (K-DBQ). The Random Forest classifier achieved 77% accuracy. Recursive feature elimination selected 10 dominant predictors; notably, engaging in non-driving-related tasks, reduced on-road gaze, and older age were significantly associated with longer takeover times. Although K-DBQ scores were not directly correlated with takeover time, their inclusion improved model robustness, consistent with ensemble learning from weak yet complementary signals. The proposed model can be integrated into advanced driver assistance systems (ADASs) to proactively identify drivers likely to exceed the 4-second takeover window, support targeted interventions, and enhance human-centered transition safety in ADSs.

**Keywords:** automated driving system; human factors; prediction modeling; takeover time; non-driving-related task; driver monitoring; ADAS



Academic Editor: Bai Li

Received: 3 December 2025
Revised: 4 March 2026
Accepted: 14 March 2026
Published: 26 March 2026



## 1. Introduction

While automated vehicles have been poised to revolutionize ground transportation systems, they also pose new challenges. One of these challenges is takeover transitions in highly ADSs [1,2]. In highly ADSs, drivers are no longer required to actively monitor the driving environment and are allowed to fully engage in non-driving-related tasks (NDRTs) [3]. However, resuming manual driving can be challenging because switching from an NDRT to manual driving requires attentional resources and a time-consuming reconfiguration of the driver's physical and cognitive state [4].

The ability of drivers to resume manual driving remains a significant road safety issue. However, there is a lack of research on the types of drivers who can safely regain vehicle control. The authors present a classification model to predict drivers requiring a

 https://doi.org/10.3390/electronics15071390

long takeover time and propose a method for applying an ADAS to improve the driver's takeover performance and mitigate this challenge. This is a convergence study that combines cognitive science, psychology, AI, and statistics in the automated vehicle domain. It collects and analyzes human factors and reaction data using elaborately designed real-world simulation experiments to provide insight into human drivers' behavior during takeover situations. Based on this understanding, we present a vision for safe automated vehicles in the future.

The expeditious establishment of international regulations supporting enhanced road safety by defining the requirements for safe takeover from the ADS to the driver is a key factor in the evolution of ADS technology and market expansion. The first United Nations (UN) Regulation for ADS, UN Regulation No. 157 [5], is a classic example. The regulation requires the ADS to provide a takeover request (TOR) (a.k.a. transition demand) and escalate it if the human driver does not resume manual driving within 4 s after the initiation of the TOR for a safe takeover. This time interval was determined by the National Traffic Safety and Environment Laboratory after a simulation study showed that most drivers could switch to manual driving without colliding if their TOR was given 4 seconds before a potential collision occurred [6].

Although 4 seconds is a critical time duration for a safe takeover, to the best of our knowledge, no current research exists that makes computational models capable of predicting which types of drivers require more than 4 seconds to regain vehicle control. Most previous studies have focused on the statistical analysis of takeover time and quality [7–17]. In the present study, therefore, we aimed to fill this research gap by developing a machine learning model that predicts drivers' takeover time and to reveal human personality and behavior traits that affect their takeover time. The Random Forest (RF) method was chosen to predict the takeover time because of its high performance and excellent generalization. We used the recursive feature elimination (RFE) algorithm to identify the most important predictors of human factors to improve the performance of the RF model.

### 1.1. Factors Influencing Takeover Time

Time-related performance is closely associated with traffic accidents that may occur during takeovers. To facilitate takeover transitions, several empirical studies have been conducted to investigate the factors that influence drivers' takeover time. Three factors have been presented as possible causes for delayed takeovers when regaining vehicle control: the characteristics of the NDRT [7–13], the situational awareness of the driver [13–15], and the driving environment [16–18].

First, several previous studies have shown that drivers performing NDRTs had longer takeover times, shorter minimum times to collision (TTCs), and more crashes in high-traffic situations than those not performing NDRTs [7–9]. The effects of NDRT modality on takeover time were also investigated. Wandtner et al. [10] and Horakova et al. [11] reported that a visual task performed on a handheld device increased takeover time and led to a higher collision rate, whereas an auditory task led to a takeover time similar to that without any task. A comprehensive meta-analysis by Hu et al. [12] synthesized these findings and confirmed that NDRTs with visual–manual interactions significantly prolong takeover time compared to other NDRT modalities. In addition, Kim et al. [13] explored the effects of visually and mentally distracting tasks and found that a visually distracting task with a high load increased takeover time more than a mentally distracting task.

Second, some studies have found that situational awareness affects takeover time. Clark et al. [14] investigated the influence of situational awareness on takeover time using a series of video clips of simulated automated driving and a standardized situational awareness measurement tool, the situation awareness global assessment technique (SAGAT) [15].

They reported that low situational awareness was associated with a longer takeover time. Kim et al. [13] confirmed that the takeover time performance improved when situational awareness information was provided to the terminal mounted in the vehicle. Even if the driver performed an NDRT, they could recognize the situational awareness information provided on the terminal. Situational awareness plays a critical role in the decision-making process across a wide range of applications and affects the takeover time during a takeover transition in an ADS.

Third, factors of the driving environment, such as the road situation, traffic density, and weather conditions, also affect takeover time. Gold et al. [16] and Du et al. [17] reported that a heavy traffic density in takeover situations led to longer takeover times and shorter times to collision. Li et al. [18] found that the takeover time in severe weather conditions was longer on highways than on city roads.

Most previous studies have focused on the effects of certain factors on takeover time. However, knowing the relationships between certain factors and takeover time is not sufficient to accurately predict drivers' takeover time in real automated vehicles because many influential factors can interact with one another. A computational model that can predict takeover time under various factors is required.

### *1.2. Existing Machine Learning Models for Takeover Time Prediction*

Recent advances in ADSs have shifted from purely descriptive analyses of factors influencing drivers' takeover time to predictive modeling, with a growing emphasis on machine learning (ML) approaches and human factor variables [19–22].

Kim et al. [19] proposed a machine learning algorithm for predicting the takeover times of drivers. The study categorized takeover time into good and slow based on an average takeover time of 4.25 s. Data were collected using a driving simulator with 90 participants. The input features were demographic information and the number of repetitions of the experiment. Using the classification and regression tree (CART) machine learning algorithm, the authors predicted the takeover time class with an accuracy of 64.3%. The study showed classification performance limitations because they only developed a model with limited driver characteristics.

To predict takeover performance, Du et al. [20] analyzed 828 takeovers using data from fixed-based driving simulator experiments and proposed an RF classifier to predict the takeover performance of drivers. Their approach to producing ground truth data for learning was to deem a takeover transition as bad performance if any of the calculated takeover times, maximum resulting accelerations, or standard deviations of the road offset values were larger than $\mu + 2\sigma$, whereas others were labeled as having good performance. Their study classified takeover performance based on the input features of physiological data, including galvanic skin response, heart rate, and eye-tracking metrics, while the driving environment data included traffic density, scenario type, and TOR lead time. The results showed that the classifier had an accuracy of 84% and an F1-score of 64%. However, there are limitations to implementing this study's system in real automated vehicles because the drivers had to wear physiological sensors to obtain physiological data.

In a recent study, Liang et al. [21] presented a CatBoost model for predicting drivers' takeover times and identified perceived spare capacity (pSC) as the most influential predictor. Their findings suggest that, rather than relying on static individual traits, the real-time monitoring of a driver's cognitive state may be a more effective approach for developing safe and personalized automated driving systems. However, although the study provides useful insights into drivers' cognitive spare capacity, it has a critical limitation: pSC is obtained through post-transition questionnaires and cannot be measured before a takeover request occurs, making it unsuitable for real-time driver monitoring in an ADS.

https://doi.org/10.3390/electronics15071390

Furthermore, Chen et al. [22] developed an XGBoost model to predict takeover time. This model integrated real-time situation awareness with individual and environmental factors. The inclusion of metrics from eye trackers significantly improved the model fit when they compared it to baseline models. SHAP analysis confirmed that situation awareness indicators, which included gaze and scan durations, were the most influential predictors of takeover performance. Their finding showing that situation awareness information from eye trackers significantly impacts takeover time aligns with the conclusions of the present study. Specifically, our RFE and statistical analysis results demonstrate that the gaze ratio on the road is a significantly important factor for the prediction of takeover time.

### *1.3. Objectiveness and Contributions of This Study*

The main objectives of this study were (1) to identify the human factors useful for predicting poorly performing drivers who require more than 4 s to regain vehicle control and (2) to predict them in advance of TOR. Thus, we selectively considered candidate human factors that were feasible to collect before TOR. As a result, we selected the type of NDRT performed during automated driving and the drivers' eye movements, which are commonly used to indicate situational awareness, as candidate human factors. Moreover, we also added drivers' demographic features and the driving behavior questionnaire (DBQ), which has been widely used as an instrument to measure risky driving behavior [23–25], as candidate human factors based on our hands-on experience and knowledge. The DBQ was originally developed to assess driving behavior in manual driving situations. However, we applied it to predict takeover time in a takeover situation because we believed that driving behavior would affect both manual driving and takeover transition.

Our study contributes to the literature in four ways: First, our study aimed to predict drivers' takeover times when they were engaged in different NDRTs. To investigate the effect of an NDRT on takeover time, we clustered NDRTs with similar visual, auditory, cognitive, and psychomotor (VACP) demands using a K-means algorithm. Subsequently, two representative NDRTs from different clusters were selected for experimental comparison, and one driving-related task (DRT), traffic environment monitoring, was selected for the comparison group and used to build the model.

Second, we carefully and comprehensively considered human personality traits and behaviors to improve the prediction results. Previous research on computational models to predict drivers' takeover time did not comprehensively consider the drivers' observable characteristics in a real automated vehicle situation.

Third, our study employed an RF model to predict takeover time and found human factors to predict poorly performing drivers using RFE combined with the RF method. To the best of our knowledge, this is the first attempt to derive human factors for predicting takeover time, reveal their importance, and propose a classifier that predicts takeover time based on UN Regulation No. 157.

Finally, we deduced insights to understand human drivers in the takeover situation through a discussion of the experimental results. Based on our knowledge and experience in cognitive science, psychology, and machine learning, we have described the effects of drivers' task-switching, situational awareness, and the selective attention shift ability of cognitive processes on takeover time. We propose an ADAS technology to reduce the risk of driver misuse and develop safe future automated vehicles based on this understanding of human drivers.

## 2. Methods

### *2.1. NDRT Clustering*

According to the multiple-resource theory, people have separated visual, auditory, cognitive, and psychomotor (VACP) attentional resources, and each VACP resource has a fixed capacity [26]. Resources can be overloaded if demand exceeds their capacity. Humans can simultaneously perform tasks with minimal interference using different resources. However, if a human attempts to perform two tasks simultaneously using the same resource, interference may occur, which can negatively affect performance. Thus, overload can occur in a single resource (visual, auditory, cognitive, or psychomotor (VACP)) or in a combination of these. If drivers resume manual driving while they are performing an NDRT, which requires high VACP demand, the demand for drivers' VACP resources may exceed their capacity. As a result, an overload can occur and deteriorate takeover performance.

To evaluate the effect of an NDRT on takeover time, we clustered the NDRTs and selected the representative NDRT. We evaluated the VACP demands for 17 NDRT candidates: sleeping, doing nothing, watching scenery, listening to music, conversation, reading, texting, browsing the Internet, using a smartphone, using a computer, drinking, putting on makeup, dressing up, working on the backseat, cleaning, watching a movie, and playing a game. Table 1 shows the VACP demands, which were assessed according to the work of Kim [27]. Sleeping required no VACP demand, whereas drinking imposed high visual and psychomotor loads, and texting required high visual, cognitive, and psychomotor loads. Correlation analysis of the evaluated VACP data showed a strong positive relationship between visual and cognitive demands ($r = 0.65$, $p = 0.004$ **). When a TOR occurs during a visually demanding NDRT, the driver's cognitive resources are more likely to be overloaded, which in turn prolongs transition time and increases the risk of traffic crashes. The regain of manual vehicle control requires high cognitive resources.

**Table 1.** VACP demands for several tasks.

| Tasks | Visual | Auditory | Cognitive | Psychomotor |
|---|---|---|---|---|
| Sleeping | 0.00 | 0.00 | 0.00 | 0.00 |
| Doing Nothing | 0.14 | 0.00 | 0.14 | 0.00 |
| Watching Scenery | 0.14 | 0.00 | 0.14 | 0.14 |
| Listening to Music | 0.14 | 0.21 | 0.14 | 0.00 |
| Conversation | 0.14 | 0.70 | 0.76 | 0.00 |
| Reading | 0.84 | 0.00 | 0.66 | 0.31 |
| Texting | 0.84 | 0.00 | 0.76 | 0.66 |
| Internet | 0.84 | 0.00 | 0.76 | 0.37 |
| Using a Smartphone | 1.00 | 0.00 | 0.76 | 0.37 |
| Using a Computer | 1.00 | 0.00 | 0.76 | 0.37 |
| Drinking | 0.71 | 0.00 | 0.14 | 1.00 |
| Putting on Makeup | 0.71 | 0.00 | 0.36 | 0.41 |
| Dressing Up | 0.57 | 0.00 | 0.17 | 0.83 |
| Working in the Backseat | 0.57 | 0.00 | 0.17 | 1.00 |
| Cleaning | 0.71 | 0.00 | 0.17 | 0.83 |
| Watching a Movie | 1.00 | 0.70 | 0.61 | 0.00 |
| Playing a Game | 1.00 | 0.70 | 0.76 | 0.37 |

Using the K-means algorithm, we clustered the NDRTs based on the input features listed in Table 1. K-means is one of the most widely used and straightforward clustering algorithms [28]. It groups NDRTs with similar VACP demands and minimizes the within-cluster sum of squared distances for an optimized number of clusters K.

To determine the optimal number of clusters, we calculated the total within-cluster distances for increasing values of K and identified the "elbow point", where further increases in K yielded only marginal reductions in distance. As shown in Figure 1, the within-cluster distance did not decrease noticeably beyond five clusters. Therefore, we set K = 5 and grouped the 17 NDRTs into five clusters, as summarized in Table 2.

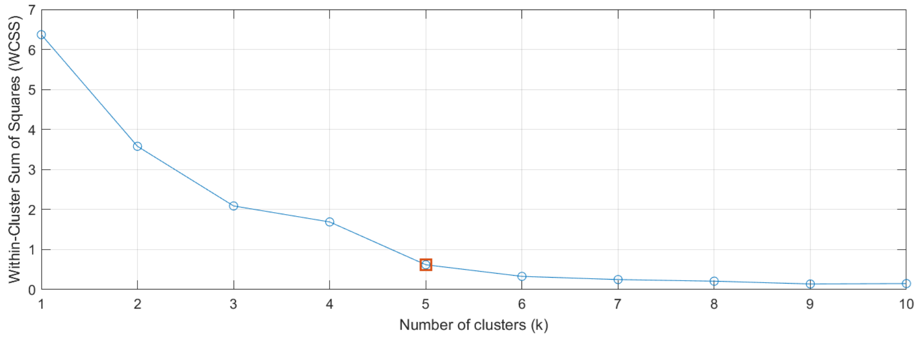


**Figure 1.** Within-cluster sum of squares; The red box highlights the selected optimal number of clusters (K = 5) based on the elbow point.

**Table 2.** VACP demands and clustering results for NDRTs.

| Tasks | Average Demand | | | |
|---|---|---|---|---|
| | **Visual** | **Auditory** | **Cognitive** | **Psychomotor** |
| Sleeping<br>Doing Nothing<br>Watching Scenery<br>Listening to Music | 0.11 | 0.05 | 0.11 | 0.04 |
| Conversation | 0.14 | 0.70 | 0.76 | 0.00 |
| Drinking<br>Putting on Makeup<br>Dressing Up<br>Cleaning<br>Working in the Backseat | 0.64 | 0 | 0.16 | 0.77 |
| Using a Smartphone<br>Using a Computer<br>Reading<br>Internet<br>Texting | 0.87 | 0 | 0.68 | 0.42 |
| Watching a Movie<br>Playing a Game | 1.00 | 0.70 | 0.69 | 0.19 |

Among the five clusters of NDRTs, this study selected two tasks for the experiment—texting and drinking—which are clearly distinguishable in terms of their visual, cognitive, and psychomotor load levels and can be safely replicated in a driving simulator environment. This selection was an experimental design decision aimed at ensuring experimental control and minimizing participant fatigue and learning effects by limiting the number of NDRTs assigned to each participant. Although we tested only two representative NDRTs, these tasks represent the extreme ends of psychomotor and cognitive load. This approach ensures that a sufficient contrast exists between NDRT types. Furthermore, our previous research [29] confirmed that these tasks significantly and distinctly degraded subjective driver readiness when we compared them to the DRT, which monitored the road environment. In particular, the texting task caused significantly more degradation in driver readiness than drinking did. Driver readiness is essential because it influences successive intervention performance when the driver must regain control of the vehicle from the system and continue to drive manually. This selection ensures that the dataset includes a variety of driver states that reflect different levels of readiness.

### 2.2. Driving Behavior Questionnaires

The DBQ is one of the most widely used subjective instruments for measuring self-reported driving behavior. Since Reason et al. [30] introduced the DBQ, many studies have assessed risky driving behaviors in a wide range of countries. Researchers have developed country-specific versions of DBQs to measure different cross-cultural driving traits [23–25]. In this study, we used a shortened version of the Korean-DBQ (K-DBQ) [31], presented in Table 3.

**Table 3.** Description of K-DBQ.

| Item No. | Questions |
|---|---|
| 1 | The more speed up, I feel the more pleasant |
| 2 | I feel stress reduced when driving at high speed. |
| 3 | I tend to drive impatiently. |
| 4 | I think that following the speed limit is the behavior of preventing traffic flow. |
| 5 | I think following the speed limit on the highway is inflexible behavior. |
| 6 | I have heard that you drove the car like you were going to have an accident. |
| 7 | When the traffic light turns green, I am leaving faster than other vehicles. |
| 8 | I think the speed limit should be higher than it is now. |
| 9 | I do not yield my lane well when driving. |
| 10 | I lack concessions when driving. |
| 11 | I get angry at other drivers who try to pass. |
| 12 | I get annoyed by the red light while driving. |
| 13 | I am frustrated by slow-moving vehicles. |
| 14 | I tend to start and stop suddenly when driving. |
| 15 | I think it's a waste of time to rest at a highway rest area because of fatigue. |
| 16 | I don't want to allow it when another car is about to pass by. |

### 2.3. Participants

A total of 46 participants were recruited from a job-recruiting website and a social networking site. They received compensation of $27 to complete the experiment. They were required to have a valid driver's license and be free of any autonomic nervous system disorders or motion sickness. There were 24 males and 22 females ranging in age from 23 to 53 years, with a mean of 38.83 years old (standard deviation (SD) = 10.12). The participants' driving experience ranged from 1 to 30 years, with a mean of 12.33 years (SD = 8.73). This research was approved by the Korea National Institute for Bioethics Policy Institutional Review Board (IRB) (Approval No. P01-202009-13-001). All participants provided written informed consent.

### 2.4. Experimental Environment

We constructed a driving simulator based on Hyundai Click with fully operational automated driving functionality (a.k.a. level 3), as shown in Figure 2. The driving simulator was equipped with a three-channel visual display composed of three monitors (43 in each) on the front, left, and right sides of the road to obtain a 135-degree horizontal field of view. The rearview and wing mirrors were also equipped for the rear road scene, with horizontal and vertical fields of view of 20 and 25 degrees, respectively. An in-vehicle agent providing situational information was installed on the center fascia of the driving simulator. Figure 3 shows the situational information provided by the in-vehicle agent: driving mode, driver readiness, TOR information, ego-vehicle speed, and traffic situation information. The TOR

https://doi.org/10.3390/electronics15071390

information includes the time left and distance until the TOR, which are presented as values and progress bars.

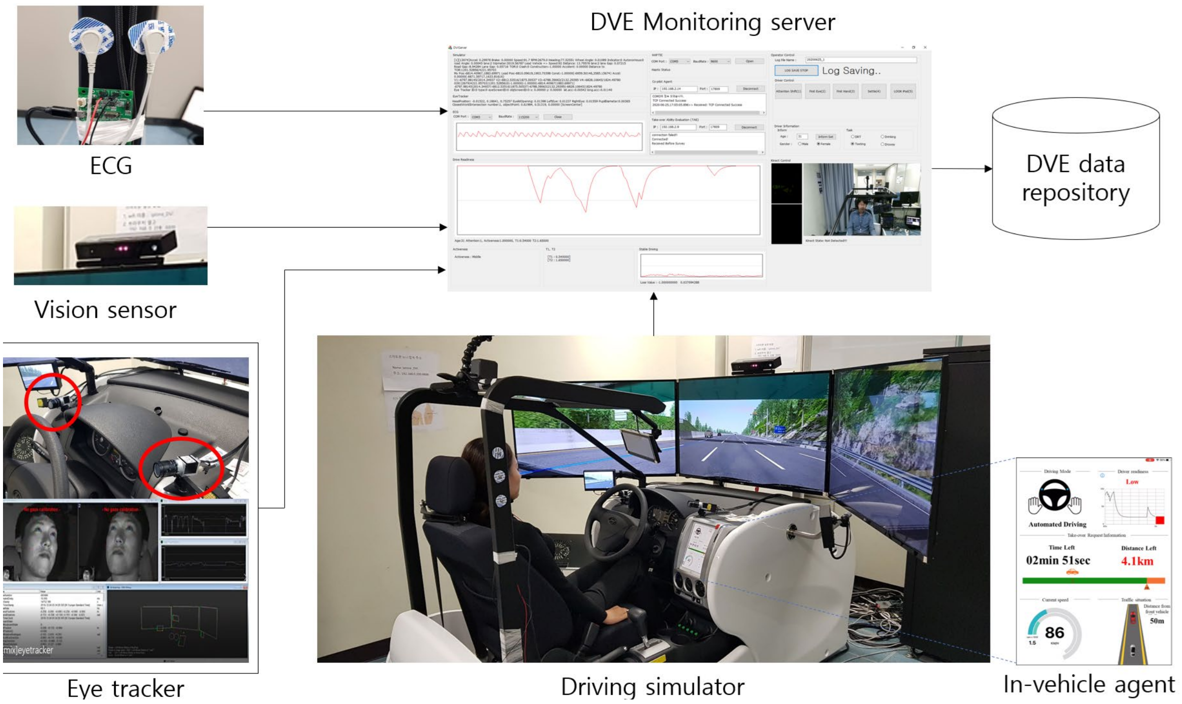


**Figure 2.** Experimental environment.

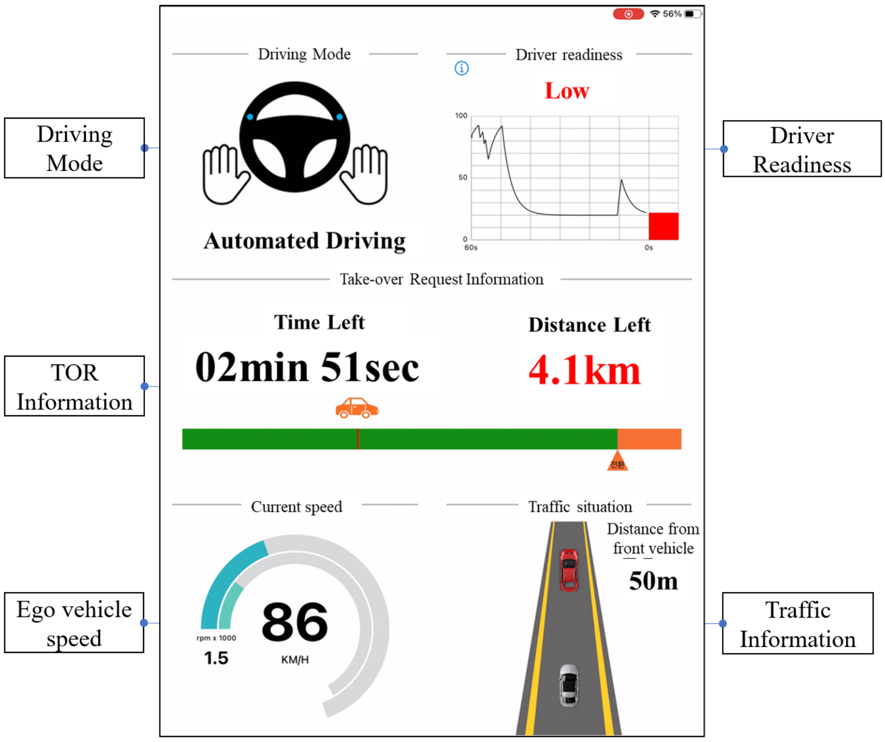


**Figure 3.** Situation information visually provided by the in-vehicle agent.

The driving simulator was connected to a vision sensor; an infrared eye-tracker, i.e., a Smart Eye Pro 8.0 [32]; an ECG sensor developed by ETRI; and a DVE (Driver, Vehicle, Environment) monitoring server that collected and synchronized the measured data to investigate the participants' reactions and behavior during the experiment, as described in

Table 4. The DVE data were collected and saved to the DVE monitoring server, enabling 5 ms high-resolution time-synchronized data acquisition.

**Table 4.** Description of experimental setup.

| Items | Description |
|---|---|
| Driving simulator | Fully operational ADS simulator based on Hyundai Click, providing (1) ego-vehicle information such as speed, RPM, distance to the vehicle ahead, lane information, and vehicle position in the lane and (2) environmental information such as weather conditions and surrounding vehicle information |
| In-vehicle agent | The device that provides situational information and the TOR alarm installed in the center fascia of the driving simulator |
| Vision sensor | The Kinect camera installed on the front monitor of the driving simulator for driver motion detection |
| Eye tracker | Two infrared cameras installed on the left and right corners of the driving simulator dashboard to measure the driver's gaze information. |
| ECG sensor | The physiological sensor device collecting a driver's heart rate and rhythm at a 200 Hz sampling rate using two disposable gelled Ag/AgCl electrodes on the chest |

The experimental road screen consisted of an eight-lane two-way highway environment with four lanes each. The test vehicle was in the third lane of the four, and the traffic density was low, that is, less than seven vehicles per kilometer (VPK), with the vehicles in front of the test vehicle during the entire scenario.

The driving simulator provided a TOR alarm through auditory and visual channels to alert the driver to resume manual driving. The TOR alarm is conveyed by a voice message as "Please start manual driving" and by visual images from the in-vehicle agent. Manual driving was activated when participants pressed the driving mode button located next to the steering wheel to change the driving mode.

### *2.5. Experimental Procedure*

Figure 4 shows the experimental procedure. Prior to the start of the experiment, sufficient practice driving and takeovers were conducted until the participants were familiar with the simulator environment, which took an average of 5 min. After the practice, the participants completed pre-questionnaires on the K-DBQ, marked on a 1–7 Likert scale, and provided demographic information. Each participant was involved in three experiments, using a within-participant design. Every experiment was initiated in automated driving mode, and the participants engaged in three different tasks (drinking water, texting on the smartphone, or monitoring traffic environments), as depicted in Figure 5. The participants often drank water while holding a water bottle in one hand during the drinking task, and they continued typing texts with a smartphone during the texting task. After two minutes of automated driving, the TOR alarm, which consisted of an audible message, "Please start manual driving", and a visual image, was generated. After the TOR, participants resumed manual driving. Post-questionnaires were conducted to determine driver readiness, NASA-TLX mental workload measures [33], and the level of situational awareness scored on a 0–10 Likert scale after finishing each experiment.

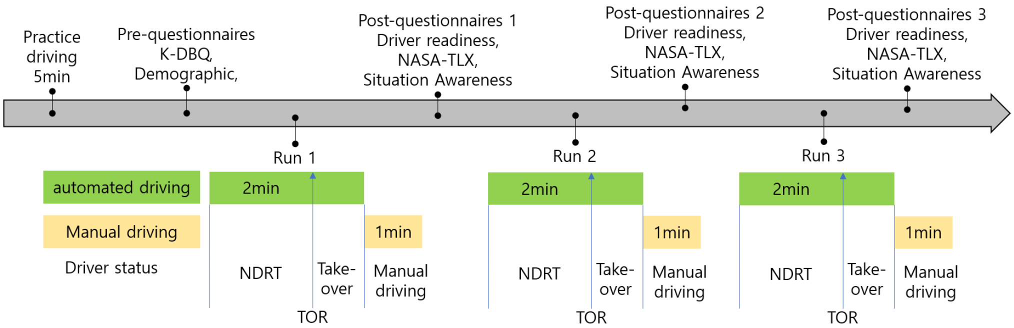


**Figure 4.** Illustration of the experimental procedure for three takeover events.

(**a**) Drinking (**b**) Traffic monitoring (**c**) Texting

**Figure 5.** NDRTs in which the drivers engaged during automated driving.

## *2.6. Data Preprocessing*

Data points from 26 of the 46 participants were used for the experimental data analysis. Fifteen data points were removed due to failure to collect eye-tracking data. Three data points were removed owing to systematic errors, and two data points were removed owing to outliers. Because each subject participated in three experiments, a total of 78 takeover experiment results were used in this study. For outlier detection, we used the outlier labeling rule proposed by Hoaglin et al. [34,35], which is based on multiplying the interquartile range (IQR) by a factor of 1.5. Based on this rule, any data points that lie beyond the (Q1 − 1.5IQR) or (Q3 + 1.5IQR) limits are considered outliers.

For transition performance analysis, we measured takeover time as the time from the TOR alarm to the pressing of the driving mode button to start manual driving. We then categorized the drivers who required longer than 4 s after TOR as the poorly performing group and the others as the good-performing group. Eventually, we obtained an imbalanced dataset with 25 "poorly performing" and 53 "good-performing" labels.

## *2.7. Model Construction and Prediction Process*

The RF method was used to classify and predict takeover time. RF is one of the most commonly used machine learning algorithms owing to its high performance and generalization. RF, developed by Breiman [36], is an ensemble machine learning model that facilitates classification by combining large sets of decision trees. This approach uses a bagging (i.e., bootstrap) operation to randomly fit a large number of decision trees to different subsamples of the dataset and employs an averaging technique to improve prediction accuracy and control overfitting [37]. In this process, multiple classification trees were generated using a random subset of the sample. The multiclassification tree was then voted for correct classification by a majority [38]. According to the literature, the advantage of RF is that it is distribution-free and does not suffer from the Hughes phenomenon of overfitting [39]. It is also reliable, fast [40], and easy to implement and interpret [41].

Figure 6 shows the classification model construction and performance evaluation process. Data preprocessing is the first and most important step in the process because it can improve the prediction accuracy. During data preprocessing, we detected and removed outliers from the raw dataset described in Table 5 and further normalized them using the StandardScaler technique to remove the mean and scale-to-unit variance. Centering and scaling were independently performed for every feature. Scaling was performed to eliminate the possible domination of features that have variance orders of magnitude greater than the other features.

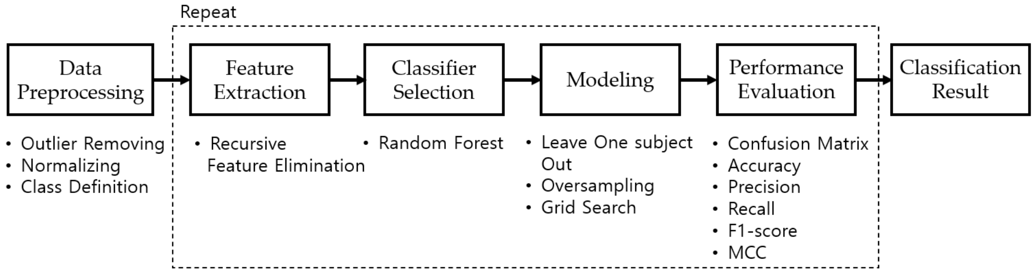


**Figure 6.** Classification model construction and performance evaluation process.

**Table 5.** Description of input dataset.

| Category | ID | Feature | Description | Value |
|---|---|---|---|---|
| Human behavior | A | NDRT type | Non-driving-related task during automated driving | M/D/T |
| | B | On-road gaze ratio | Time ratio of gazing at the road during total automated driving period | 0–100 |
| | D | K-DBQ-1 | Self-reported score on K-DBQ-1 | 1–7 |
| | I | K-DBQ-2 | Self-reported score on K-DBQ-2” | 1–7 |
| | P | K-DBQ-3 | Self-reported score on K-DBQ-3 | 1–7 |
| | N | K-DBQ-4 | Self-reported score on K-DBQ-4 | 1–7 |
| | Q | K-DBQ-5 | Self-reported score on K-DBQ-5 | 1–7 |
| | T | K-DBQ-6 | Self-reported score on K-DBQ-6 | 1–7 |
| | J | K-DBQ-7 | Self-reported score on K-DBQ-7 | 1–7 |
| | G | K-DBQ-8 | Self-reported score on K-DBQ-8 | 1–7 |
| | S | K-DBQ-9 | Self-reported score on K-DBQ-9 | 1–7 |
| | O | K-DBQ-10 | Self-reported score on K-DBQ-10 | 1–7 |
| | H | K-DBQ-11 | Self-reported score on K-DBQ-11 | 1–7 |
| | F | K-DBQ-12 | Self-reported score on K-DBQ-12 | 1–7 |
| | K | K-DBQ-13 | Self-reported score on K-DBQ-13 | 1–7 |
| | L | K-DBQ-14 | Self-reported score on K-DBQ-14 | 1–7 |
| | M | K-DBQ-15 | Self-reported score on K-DBQ-15 | 1–7 |
| | R | K-DBQ-16 | Self-reported score on K-DBQ-16 | 1–7 |
| Demographic information | C | Age | Driver's age | 20–60 |
| | E | Driving experience | Cumulative number of driving years | 2–30 |
| | U | gender | Driver's gender | M/F |

In machine learning, feature selection algorithms are used to improve the performance of models and reduce their training time. The most accurate way to select feature subsets is to investigate all possible combinations of features, that is, use an exhaustive grid search methodology. However, this is impractical because it takes too much time to investigate all combinations of features by training the model and testing it with each subset. In our experiment, the dataset was composed of 21 features, and we had 221 = 2,097,152 possible feature subset combinations for evaluation. If our machine learning model took just one minute in the training and testing phases, then it would have taken 2,097,151 minutes or 1456.36 days, which would not have been feasible. Therefore, we used the recursive feature elimination (RFE) algorithm proposed by Guyon et al. [42] for feature selection. It has been used in gene microarrays, which have thousands of features. The RFE iteratively eliminates the least important feature and performs classification based on the new feature subset. All the feature subsets were evaluated based on their classification performance. The authors used the RFE method in a support vector machine (SVM), which also works in conjunction with other classification methods. In this study, we combined the RFE with an RF classifier. The features were ranked and scored according to their importance using the Gini importance ranking method. Algorithm 1 shows the pseudocode for the Random Forest combined with RFE (RF-RFE). RF-RFE calculates the importance scores of the features, according to which it sorts them in descending order, and removes the least important features from the feature set. Before removing one feature at a time from the previous feature set, the prediction accuracy was evaluated. Recursive feature elimination was repeated until one feature remained. The feature subset that provided the best precision accuracy was the optimal feature.

**Algorithm 1** RF-RFE Algorithm

```
F: = Set of features = { f_1, f_2, ..., f_p }
B: = Set of features proving best accuracy, the subset of whole features
S: = The score of feature importance for a particular feature set using Gini importance
ACC: = the classification accuracy with a particular feature set
Input:
   Set of p original features, F = {f_1, f_2, ..., f_p}
   ACC_best = 0
Output:
   Subset of features B providing best accuracy
Begin
   Repeat for i in (1:p − 1)
       Train the Random Forest model and measure the classification accuracy
       ACC with a particular feature set F
       if (ACC >= ACC_best )
           ACC_best = ACC
           B = F
       end if
       Calculate score of feature importance for a particular feature set F
       S = {(f_i, s_i) | f_i in F}
       Rank the S
       f_last ← last ranked feature in S
       F ← F − f_last
   End Repeat
End
```

In this study, the number of participants in the good-performing group was dominant. This resulted in an imbalance between the good and poor groups. To address this data imbalance problem, we performed data augmentation for the minority group of poorly performing drivers. This is an effective method for solving the common problem of imbalanced classes in machine learning classification and has been widely accepted in both academia and industry [43,44]. The synthetic minority oversampling technique

(SMOTE) [44] was used as the oversampling algorithm, which synthetically increases the minority class based on k-nearest neighbors to balance the dataset.

To ensure the model's generalizability and prevent data leakage between trials from the same participant, we utilized the Leave-One-Subject-Out Cross-Validation (LOSO-CV) method for model selection and evaluation. LOSO-CV provides a rigorous criterion by ensuring that the model is tested on an entirely unseen subject, which is a sensible approach for evaluating the performance of machine learning models in human factor research [45–47]. Specifically, instead of selecting a single trial, we iteratively selected all trials belonging to one subject (out of 26 participants) as the test set. The experimental data from the remaining subjects were then oversampled using SMOTE to resolve the imbalance and used to train the RF classifier.

During the training phase, we utilized an exhaustive grid search strategy to optimize the hyperparameters. We focused on two key parameters: the number of decision trees (estimators) and the maximum depth of each tree [48]. For the number of estimators, we selected values of 100, 200, 300, 400, and 500. We tested the maximum depth at levels of 5, 10, and 20. We expected that the RF model, combined with hyperparameter tuning, RFE, and the robust LOSO-CV framework, would provide a more reliable and generalized prediction of takeover performance.

We then comprehensively evaluated the constructed prediction model using the commonly used accuracy, F1-score (weighted harmonic mean of precision and recall), precision, recall, [49], and Matthews correlation coefficient (MCC) [50].

## 3. Results

### *3.1. RF-RFE Method for Predicting Poorly Performing Drivers*

We used the RF-RFE method to predict takeover time as a classification problem. We evaluated the performance of the RF-RFE method, which classifies and predicts poorly performing drivers. The RFE produced feature subsets of different sizes ranging from 21 to 1, where 21 denoted the full input feature size in the dataset. For each feature subset, the hyperparameter-optimized version of the RF classifier was obtained using the grid search algorithm. Table 6 lists the optimized RF model for each subset of features and the obtained performance. We found that the RF model had the best performance ($RF_{best}$). The model was composed of 100 decision trees with a depth of 5 and included 10 features selected by the RFE algorithm. Note that, to obtain the best performance, we sequentially removed one feature at a time from the least important one to the most important one. The $RF_{best}$ model using only 10 features showed a prediction accuracy of 77%, which was 10% higher than that of the RF model using all features ($RF_{all}$). The $RF_{best}$ model had a precision of 0.63, recall of 0.68, F1-score of 0.65, and MCC of 0.48, whereas the $RF_{all}$ model had a precision of 0.48, recall of 0.44, F1-score of 0.46, and MCC of 0.22.

We compared the importance of the features used in the $RF_{best}$ model to understand the effects of features on the model. The RF produced a feature importance measure that directly expressed the role of a feature in all interactions utilized in the model. The feature–importance plot for the $RF_{best}$ model is shown in Figure 7. It shows that the most important feature for takeover time prediction was an NDRT, which drivers performed during automated driving (53.1%), followed by the ratio of gazing at the road or at the in-vehicle agent (8.1%), driver age (6.9%), K-DBQ-1 (5.7%), driving experience (5.6%), K-DBQ-2 (4.5%), K-DBQ-8 (4.2%), K-DBQ-12 (4.1%), K-DBQ-11 (4.0%), and K-DBQ-7 (3.7%).

**Table 6.** Performance measures.

| N Features | Feature ID | N Estimators | Max Depth | Accuracy | Precision | Recall | F1-Score | MCC | Description |
|---|---|---|---|---|---|---|---|---|---|
| ***21*** | ***A–U*** | **200** | **5** | **0.67** | **0.48** | **0.44** | **0.46** | **0.22** | ***RF<sub>all</sub>*** |
| 20 | A–T | 500 | 10 | 0.59 | 0.33 | 0.28 | 0.30 | 0.02 | - |
| 19 | A–S | 200 | 10 | 0.71 | 0.55 | 0.48 | 0.51 | 0.30 | - |
| 18 | A–R | 400 | 5 | 0.65 | 0.46 | 0.44 | 0.45 | 0.20 | - |
| 17 | A–Q | 100 | 5 | 0.69 | 0.52 | 0.52 | 0.52 | 0.29 | - |
| 16 | A–P | 400 | 5 | 0.69 | 0.52 | 0.48 | 0.50 | 0.28 | - |
| 15 | A–O | 500 | 5 | 0.67 | 0.48 | 0.52 | 0.50 | 0.25 | - |
| 14 | A–N | 400 | 10 | 0.69 | 0.52 | 0.52 | 0.52 | 0.29 | - |
| ***13*** | ***A–M*** | 200 | 5 | 0.76 | 0.63 | 0.60 | 0.61 | 0.43 | |
| 12 | A–L | 300 | 10 | 0.71 | 0.54 | 0.52 | 0.53 | 0.32 | - |
| 11 | A–K | 500 | 10 | 0.74 | 0.61 | 0.56 | 0.58 | 0.40 | - |
| **10** | **A–J** | **100** | **5** | **0.77** | **0.63** | **0.68** | **0.65** | **0.48** | $RF_{best}$ |
| 9 | A–I | 200 | 5 | 0.73 | 0.58 | 0.56 | 0.57 | 0.38 | - |
| 8 | A–H | 500 | 5 | 0.69 | 0.52 | 0.52 | 0.52 | 0.29 | - |
| 7 | A–G | 200 | 10 | 0.69 | 0.52 | 0.56 | 0.54 | 0.31 | - |
| 6 | A–F | 100 | 10 | 0.69 | 0.52 | 0.56 | 0.54 | 0.31 | - |
| 5 | A–E | 100 | 10 | 0.74 | 0.62 | 0.52 | 0.57 | 0.39 | - |
| 4 | A–D | 200 | 10 | 0.76 | 0.64 | 0.64 | 0.64 | 0.47 | - |
| ***3*** | ***A–C*** | **500** | **5** | **0.64** | **0.45** | **0.56** | **0.50** | **0.23** | ***RF<sub>stat</sub>*** |
| 2 | A–B | 100 | 5 | 0.60 | 0.41 | 0.52 | 0.46 | 0.15 | - |
| 1 | A | 100 | 5 | 0.62 | 0.45 | 0.92 | 0.61 | 0.38 | - |

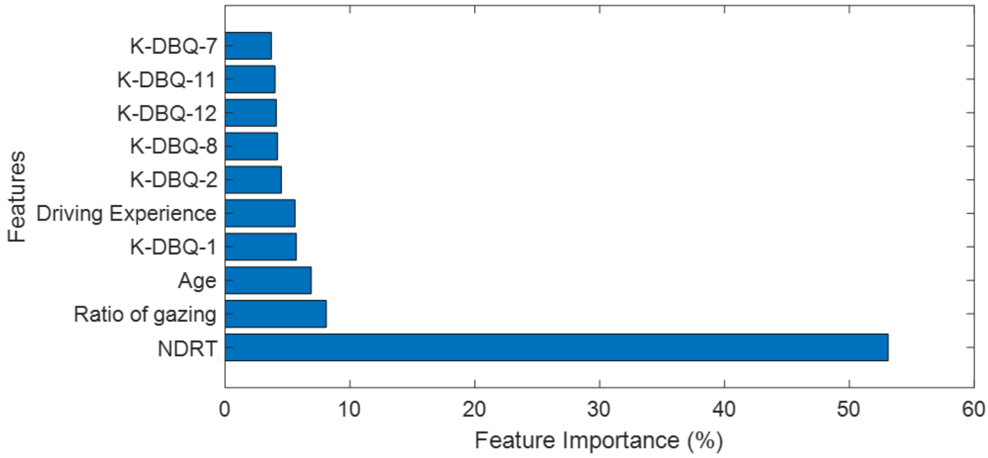


**Figure 7.** Feature importance. Ratio of gazing: ratio of gazing at the road or at the in-vehicle agent.

We found that the good- and poorly performing drivers showed statistically significant differences in the top 3 out of 10 features ($p$-value $\leq 0.05$ in the statistical analysis): (1) type of NDRT, (2) ratio of gazing at the road or at the in-vehicle agent, and (3) driver age. The other seven features showed no statistical difference between good- and poorly performing drivers. The results of the statistical analysis are described in the next subsection. In Table 6, we compare the prediction performance of the RF model ($RF_{stat}$) using only three statistically significant features with that of the $RF_{best}$ model. The $RF_{stat}$ model showed a 13% takeover time prediction accuracy degradation compared with the $RF_{best}$ model. The $RF_{stat}$ model had an accuracy of 64%, precision of 0.45, recall of 0.56, F1-score of 0.50, and MCC of 0.23.

We also determined receiver operating characteristic (ROC) curves [51] for the three types of experimental settings, that is, the $RF_{best}$, $RF_{all}$, and $RF_{stat}$ models, as shown in Figure 8. The area under the curve (AUC) is an important parameter in the ROC plot; a model with a higher AUC is considered better. The AUC for the $RF_{best}$ model was 0.72, that for the $RF_{all}$ model was 0.61, and that for the $RF_{stat}$ model was 0.61. This was consistent with the performance evaluation results presented in Table 6.

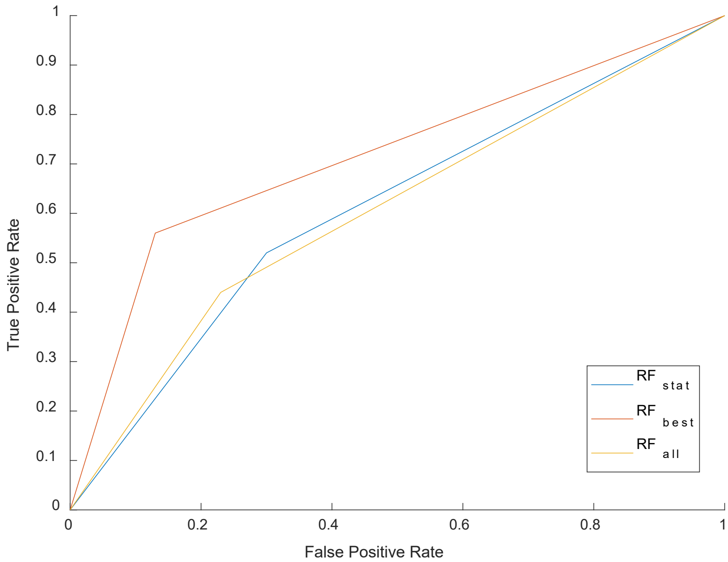


**Figure 8.** ROC curve.

### 3.2. Comparative Performance and Model Selection Rationale

To validate the selection of Random Forest (RF) as the primary classifier for identifying poorly performing drivers, a benchmark comparison was conducted against four other machine learning models—CatBoost, XGBoost, Logistic Regression, and SVM—using the full set of 21 features. The performance of each model, evaluated through Leave-One-Subject-Out Cross-Validation (LOSO-CV) to ensure generalizability across different participants, is summarized in Table 7. The results indicate that the gradient boosting models, CatBoost and XGBoost, achieved an accuracy of 0.68, which was marginally higher (0.01) than that of the RF model (0.67). However, RF significantly outperformed traditional linear models, such as Logistic Regression (0.62) and SVM (0.58), confirming its superior ability to handle the complexities of the dataset.

**Table 7.** Model comparison.

| Model | Accuracy | Precision | Recall | F1-Score | MCC |
|---|---|---|---|---|---|
| CatBoost | 0.68 | 0.50 | 0.52 | 0.51 | 0.27 |
| XGBoost | 0.68 | 0.50 | 0.48 | 0.49 | 0.26 |
| Logistic Regression | 0.62 | 0.41 | 0.48 | 0.44 | 0.15 |
| SVM | 0.58 | 0.33 | 0.32 | 0.33 | 0.02 |

Despite the marginal performance lead of boosting frameworks, RF was selected as the primary model due to its superior stability in handling human factor datasets with limited sample sizes [52]. For small training sets, boosting methods that sequentially minimize bias are often susceptible to overfitting; in contrast, RF's bagging operation averages results from multiple independent decision trees to effectively reduce model variance and enhance stability. This stability was critical for ensuring that the model remained robust when generalizing to unseen subjects within the LOSO-CV framework.

Furthermore, the significant performance gap between RF and linear models, such as Logistic Regression and SVM, confirmed that the relationships among human factors—including age, gaze behavior, and NDRT type—are inherently nonlinear and complex. RF's ensemble structure and high interpretability via Gini importance were crucial for identifying the dominant human factors required to design an effective, human-centered ADAS.

### 3.3. Statistical Analysis Results

To explain the machine learning model, we performed a statistical analysis. The two-sample *t*-test was used to investigate statistical differences in interval or ratio-scale data. The statistical difference in mean takeover time between good and poor drivers was investigated according to age, driving experience, K-DBQ score, and the ratio of gazing at the road or in-vehicle agent using a two-sample *t*-test. Furthermore, the frequencies and percentages for categorical data were computed, and the chi-square test was used to statistically analyze categorical data and the association between takeover time performance and sex.

#### 3.3.1. Takeover Time

Table 8 presents descriptive statistics associated with takeover time. The overall average takeover time was 3.65 s (SE = 0.14), the minimum was 1.17 s, and the maximum was 7.06 s. Participants in 25 of the 78 experiments were classified into the poor group, whereas 53 were classified into the good group. In other words, 32% of the drivers needed longer than 4 s to regain vehicle control. The results showed that NDRT type had a significant effect on takeover time. The poorly performing drivers needed a 70% longer takeover time (mean = 5.06 s, SE = 0.17) than the good-performing drivers (mean = 2.98 s, SE = 0.10) (t(1,77) = 10.94, $p$-value < 0.000) on average.

**Table 8.** Mean takeover time with standard error for good- and poorly performing groups.

| Item | Group | | | *p*-Value |
|---|---|---|---|---|
| | Overall | Good | Poor | |
| n | 78(100%) | 53(68%) | 25(32%) | - |
| Takeover Time (sec) | $3.65 \pm 0.14$ | $2.98 \pm 0.10$ | $5.06 \pm 0.17$ | <0.000 *** |

Values are presented as frequencies (%) or means ± SEs (Standard Errors); Sig. at ***: $p$-value < 0.001.

#### 3.3.2. Drivers' Behavior

In this study, we investigated the effect of driver behavior on takeover time. The drivers' behavior was analyzed in terms of NDRT type, the ratio of gazing at the road or at the in-vehicle agent, and the driving behavior traits measured by the K-DBQ.

First, we analyzed the takeover time according to NDRTs in which the driver was engaged during automated driving. A one-way ANOVA was conducted to analyze the effect of the NDRT on the takeover time. The results showed that the NDRT had a statistically significant effect on takeover time (F(2,77) = 12.76, $p$-value < 0.000 ***), as depicted in Figure 9. To compare the takeover time in three different NDRTs, Tukey's test was conducted, and the results showed that drivers who monitored the traffic environment had a significantly faster takeover time (mean = 2.76 s, Standard Error (SE) = 0.14) than those who were drinking (mean = 4.04 s, SE = 0.22, $p$-value < 0.000 ***) or texting (mean = 4.13 s, SE =0.26, $p$-value < 0.000 ***). However, there was no statistical difference in takeover time when the driver engaged in drinking or texting ($p$-value = 0.958). This analysis result means that drivers engaged in the DRT, traffic monitoring, resumed manual driving faster than those who were engaged in an NDRT, i.e., drinking or texting.

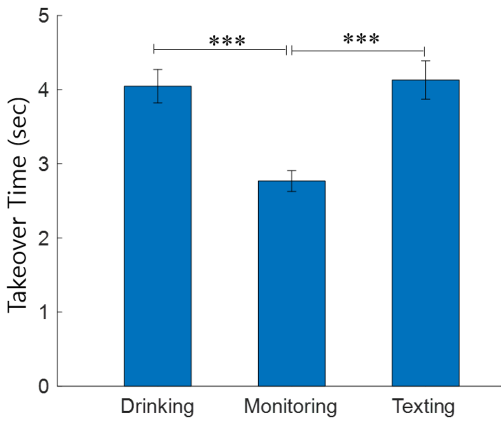


**Figure 9.** Mean takeover times with standard errors as error bars for each NDRT group. $n_{drinking}$ = 26; $n_{texting}$ = 26; $n_{monitoring}$ = 26; Sig. at *** $p$-value <0.001.

Second, we found that takeover time was negatively correlated with the ratio of gazing at the road or in-vehicle agent (r = −0.35, $p$-value = 0.001 **). A significantly longer takeover time was required when drivers gazed at the road or agent less. Figure 10 shows that poorly performing drivers gazed at the road or agent 14.23% less (mean = 32.59%, SE = 6.38) than good-performing drivers (mean =46.82%, SE = 3.98) (t(1, 77) = 1.96, $p$-value = 0.050).

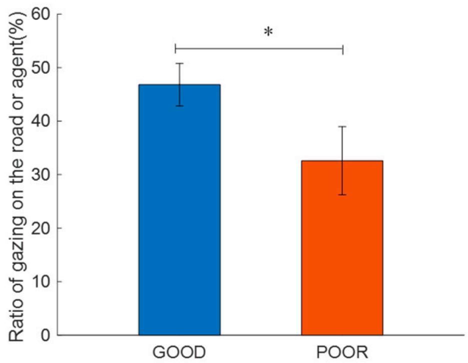


**Figure 10.** Mean ratio of gazing at the road or at the in-vehicle agent with standard errors as error bars for good- and poorly performing groups. $n_{good}$ = 53; $n_{poor}$ = 25; at * $p$-value < 0.05.

Third, we found that poorly performing drivers tended to be more frustrated with slow-moving vehicles than good-performing drivers (t(1, 77) = 3.34, $p$-value = 0.071), as shown in K-DBQ-13 of Table 9. However, good- and poorly performing drivers showed no statistical differences in any of the K-DBQ item scores.

### 3.3.3. Demographic Feature

We analyzed the influence of drivers' demographic features on takeover time, including age, driving experience, and sex. We found that takeover time was positively correlated with driver age (r = 0.25, $p$-value = 0.025 *). The poorly performing drivers were 4.99 years older (mean = 39.24 years old, SE = 2.12) than the good-performing drivers (mean = 34.25 years old, SE = 1.28) (t(1,77) = 2.11, $p$-value = 0.038 *), as shown in Figure 11.

**Table 9.** K-DBQ statistical analysis results.

| Item | Scores in Good Perf. Group | Scores in Poorly Perf. Group | *p*-Value |
|---|---|---|---|
| K-DBQ-1 | 4.47 ± 1.15 | 4.68 ± 1.77 | 0.536 |
| K-DBQ-2 | 3.75 ± 1.65 | 4.04 ± 2.05 | 0.513 |
| K-DBQ-3 | 3.96 ± 1.56 | 3.84 ± 1.49 | 0.745 |
| K-DBQ-4 | 2.89 ± 1.59 | 2.88 ± 1.62 | 0.986 |
| K-DBQ-5 | 4.57 ± 1.74 | 4.60 ± 1.94 | 0.938 |
| K-DBQ-6 | 2.66 ± 1.51 | 2.64 ± 1.44 | 0.955 |
| K-DBQ-7 | 3.77 ± 1.73 | 4.00 ± 1.91 | 0.604 |
| K-DBQ-8 | 3.85 ± 1.94 | 4.56 ± 1.94 | 0.134 |
| K-DBQ-9 | 2.94 ± 1.56 | 2.88 ± 1.56 | 0.868 |
| K-DBQ-10 | 3.23 ± 1.44 | 3.00 ± 1.47 | 0.521 |
| K-DBQ-11 | 3.33 ± 1.62 | 3.48 ± 1.90 | 0.736 |
| K-DBQ-12 | 3.64 ± 1.58 | 3.68 ± 1.82 | 0.924 |
| K-DBQ-13 | 4.07 ± 1.36 | 4.68 ± 1.38 | 0.071 |
| K-DBQ-14 | 3.13 ± 1.52 | 3.08 ± 1.68 | 0.892 |
| K-DBQ-15 | 1.89 ± 1.38 | 2.00 ± 1.16 | 0.724 |
| K-DBQ-16 | 2.47 ± 1.22 | 2.68 ± 1.35 | 0.497 |

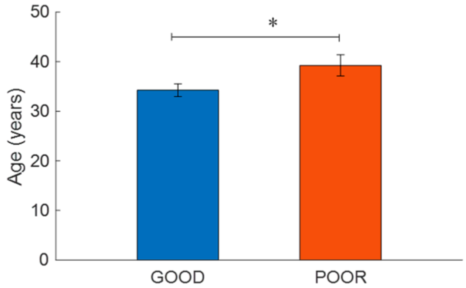


**Figure 11.** Mean ages with standard errors as error bars for each group. $n_{good}$ = 53; $n_{poor}$= 25; Sig. at *: *p*-value < 0.05.

However, driving experience was not correlated with takeover time (*p*-value = 0.180). The poorly performing drivers had an average of 13.2 years of driving experience (SE = 1.74), and the good-performing drivers had 10.1 years (SE = 1.10), but there was no statistical difference (t(1,77) = 1.53, *p*-value = 0.129), as depicted in Figure 12.

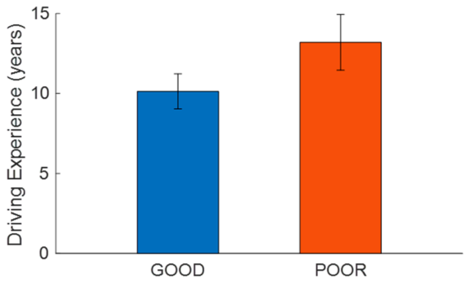


**Figure 12.** Mean driving experience with standard errors as error bars for each group. $n_{good}$ = 53; $n_{poor}$ = 25.

In addition, sex was not associated with takeover time performance ($\chi^2$ (1, N = 78) = 0.08, *p*-value = 0.777), as described in Table 10. Of the male drivers, 33.3% were classified into the poorly performing group, and 30.3% of the female drivers were classified into the poorly performing group. There was no statistically significant difference in takeover performance between sexes.

**Table 10.** Observed gender frequency for each group.

| Gender | Group | | Total |
|---|---|---|---|
| | Good | Poor | |
| Male | 30 (66.7) | 15 (33.3) | 45 (100) |
| Female | 23 (69.7) | 10 (30.3) | 33 (100) |
| total | 53 (67.9) | 25 (32.1) | 78 (100) |

Values are presented as frequencies (%).

## 4. Discussion

### *4.1. Predictive Performance and Modeling Insights*

In this study, we aimed to predict the types of drivers who require long takeover times. Thus, we propose a machine learning model for predicting drivers' takeover times using only human factors that are feasible to obtain prior to TOR. Using simulator data collected under three contrasted task conditions—texting, drinking, and traffic monitoring—the RF model achieved 77% accuracy. RFE selected 10 predictors from 21 candidate human factors, improving accuracy by 10% relative to using all features and reducing the model complexity to aid interpretability and deployment. RF is an ensemble model with N decision trees, and every tree participates in the selection of the most prominent class with f input features. Every decision tree randomly samples f features from the full F features and then uses them as the input features. Thus, some decision trees that select unimportant features can exhibit performance impairment. Therefore, we proposed the selection of optimal features using a feature selection algorithm, and then developed a machine learning model using the selected features rather than using all features.

Incorporating driving-experience indicators and selected K-DBQ scores improved the accuracy of the best RF classifier by 13% compared to the model using only the three statistically significant factors. Although the K-DBQ scores did not show statistically significant correlations with takeover time, the inclusion of these items improved the Random Forest model's accuracy. This discrepancy highlighted the inherent advantage of ensemble learning over traditional linear statistical analysis; while correlation coefficients measure isolated linear relationships, the Random Forest model captured high-dimensional nonlinear interactions between features. Specifically, the K-DBQ variables provided weak but complementary information that interacted with other features such as age, gaze ratio, and NDRT type. In ensemble learning, such weak predictors can still enhance the decision boundary and improve generalization by capturing latent behavioral tendencies that are not directly measurable. Therefore, while K-DBQ alone cannot explain takeover performance, it still contributes to the overall predictive capacity of the model through nonlinear interactions.

This study connects cognitive science with ADAS engineering through a predictive modeling framework aligned with UN Regulation No. 157, thereby positioning the work at the intersection of human factor research and safety-oriented AI deployment. Rather than maximizing predictive performance through complex architectures, our methodological focus was on developing a robust model suitable for regulatory and real-world integration. To this end, we adopted the LOSO-CV scheme, which strictly prevents data leakage and evaluates generalization to unseen drivers. This design reflects the practical deployment scenario in which the system must predict takeover performance for new users.

We selected RF as a variance-reducing ensemble method that provides greater stability than boosting approaches in relatively small datasets. While boosting models can achieve high accuracy, they are prone to overfitting when the sample size is limited. In contrast, the bagging mechanism of RF enhances robustness by averaging across decorrelated trees.

Additionally, the use of RFE supports practical system deployment by identifying the optimal set of features that maximizes predictive accuracy while simultaneously reducing feature dimensionality. Thus, the methodological contribution of this study lies not only in classification accuracy but also in demonstrating a regulation-aware, human-centered predictive framework that balances robustness, generalizability, and practical applicability in ADS contexts.

The objective of the RF model proposed in this study is to predict drivers who are likely to exhibit delayed takeover performance rather than to infer causal relationships. Delayed takeover responses can increase the risk of collisions and, therefore, pose a significant threat to road safety. Parametric econometric models [53] are well suited for causal inference under well-defined statistical assumptions. In contrast, ensemble learning approaches such as RF prioritize predictive generalization and the ability to capture nonlinear interactions among variables without imposing strict distributional assumptions. In this sense, the RF model should be interpreted as a deployment-oriented predictive tool. Meanwhile, econometric approaches remain valuable for future research aimed at quantifying the causal mechanisms underlying takeover performance. In regulatory contexts such as UN Regulation No. 157, the operational requirement is the early identification of drivers at risk of delayed takeover rather than the estimation of causal effect sizes. Therefore, predictive models that can be reliably applied to previously unseen drivers can provide practical value for safety-oriented ADAS deployment.

### 4.2. Analysis of Dominant Human Factors

In this study, we identified three dominant determinants of takeover time: NDRT type, ratio of gazing at the road or agent, and driver's age. First, the most influential feature was the NDRT type. The findings are consistent with multiple-resource theory. NDRTs that imposed visual, cognitive, and psychomotor load competed with the resources required for rapid reorientation and motor re-engagement, thereby prolonging takeover time. Additionally, this slowdown aligns with the well-known task-switch cost—an increase in reaction time when switching to a new task versus repeating the same task [54–57]. Two canonical mechanisms account for this cost: (1) task reconfiguration, an active, top–down process that establishes a new task set, and (2) residual interference from the previous task [58]. Consistent with these accounts, we observed significantly degraded takeover performance when drivers switched from NDRTs to manual driving compared with switching from a driving-related task (DRT) such as traffic monitoring. Transitioning from texting requires reconstructing driving-relevant information in working memory, and transitioning from drinking requires reconfiguring body posture and motor control under high psychomotor demand. Residual interference from the preceding NDRT can further suppress takeover performance even after reconfiguration. By contrast, traffic monitoring preserves driving-relevant working memory and posture, facilitating faster re-engagement. From a complementary perspective, Kim et al. [29] reported that lower driver readiness leads to a longer takeover time.

Second, takeover time was negatively correlated with the ratio of gazing at the road or in-vehicle agent. Drivers who looked less at the road or agent required significantly longer takeover time. We believe that the gazing ratio affected the situational awareness of the driver. Drivers usually recognize the driving situation using visual information, and thereby, gazing at the road or agent improves the driver's situational awareness. In this investigation, we found that the gazing ratio and level of TOR prediction had a positive correlation ($r = 0.35$, $p$-value = 0.002). Good-performing drivers predicted a significantly better TOR than poorly performing drivers ($t(1, 77) = 2.81$, $p$-value = 0.006). As shown in

Figure 13, the in-vehicle agent provided the time left and distance information until the TOR and helped the drivers predict the TOR in advance if they watched it.

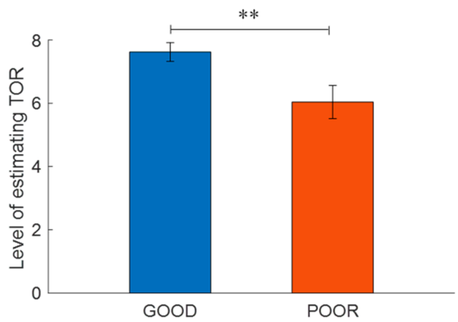


**Figure 13.** Self-reported scores of estimating TOR in advance with standard errors as error bars for each group. $n_{good} = 53$; $n_{poor} = 25$; Sig. at ** $p$-value < 0.01.

Third, driver age was more influential than driving experience, which is different from the popular belief that an experienced driver will have better driving ability. A previous study found that driving errors increase with age and are associated with degraded selective attention shift speeds, particularly when tasks require task switching and visual discrimination [59]. This was confirmed in the present study. As drivers age, their selective attention shift ability is impaired, and therefore, they require longer takeover times. As a consequence of the demographic changes in society, the number of older drivers will inevitably increase. The results of this investigation suggested that advanced driver assistance systems should be strengthened for drivers in higher age groups using ADSs. In particular, 44 years was identified as the most critical age that significantly divided the drivers into younger and older groups. Figure 14 illustrates the $p$-values determined using the $t$-test for the takeover time between the younger and older group at one-year intervals. We found the lowest $p$-value at 44 years of age.

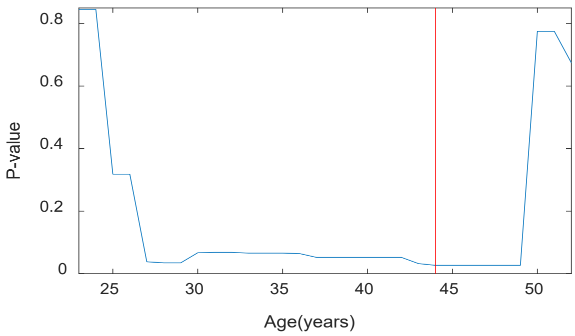


**Figure 14.** Evaluation of $p$-value of the $t$-test for the takeover time between the younger and older groups at one-year intervals.

Fourth, there is a gender stereotype that female drivers have worse driving ability than male drivers. However, we did not find any statistical difference in takeover time between sexes in this study. Furthermore, the optimal RF classifier did not use sex as the input feature to classify takeover time. It is desirable to avoid social prejudice against female drivers using ADSs.

This study has some limitations that motivate future research. First, to establish external validity, the proposed model should be replicated with larger and more diverse participants spanning a wider age range, and its performance should be rigorously evaluated in on-road situations. Second, the scope of NDRTs needs to be expanded to systematically analyze their effects on takeover time.

### *4.3. Limitations and Future Research*

While this study established a foundational RF model to predict driver takeover time, several areas for further refinement were recognized to enhance its broad applicability. Notably, although the model was constructed using high-fidelity data derived from elaborately designed simulator experiments, the findings were based on a relatively constrained dataset of 78 takeover trials from 26 participants whose ages ranged from 23 to 53 years. This constraint might influence the stability of the model. Thus, we selected the RF model utilizing the bagging method. This method provides more stability for small datasets than models that utilize boosting. We also applied SMOTE and LOSO-CV to provide a robust evaluation of the model performance. Future research needs to involve a larger number of age-diverse participants, specifically including senior drivers aged 65 and over, to validate these findings and ensure the model's generalizability across the entire lifespan.

Additionally, the selection of texting and drinking as representative NDRTs provided a clear contrast in visual, cognitive, and psychomotor loads but did not encompass the full spectrum of diverse NDRTs drivers may engage in during automated driving. Furthermore, the 4-second takeover threshold was specifically adopted from UN Regulation No. 157 as a standardized regulatory benchmark; this should be viewed as a technical reference point rather than an absolute safety limit applicable to all conditions. Future investigations should explore the impact of varying time-to-collision thresholds and establish more granular causal mechanisms to further enhance the robustness of real-time ADAS interventions. Finally, future studies should test the model in urban areas and in scenarios where traffic is heavy and complex so that the results reflect real-world driving conditions.

## 5. Conclusions

A successful takeover from an automated vehicle to a human driver is a critical safety issue in ADSs. Serious safety problems may occur if drivers fail to take over within the available time budget. In this study, we collected and analyzed the human factors affecting takeover time through elaborately designed simulation experiments. We found that 32% of the drivers required more than 4 s to regain manual vehicle control. We note that it is desirable to resume manual driving in the last 4 s after TOR for road safety, according to UN Regulation No. 157 [5]. Thus, in order to improve road safety, it is essential to identify drivers who require longer to take over the vehicle before the TOR alarm and to prepare them for vehicle re-engagement.

This study developed an RF model to predict driver takeover time in a highly automated driving system. The RF classifier achieved an accuracy of 77%. Additionally, we identified 10 human factors for takeover time prediction using the RFE algorithm. Among these, the top three important factors showed a statistically significant association with takeover time. Engagement in the NDRT, demanding high psychomotor and cognitive resources and less frequent gazing at traffic and situational information during automated driving, was associated with slower takeover times. In addition, older age was also associated with longer takeover time.

Although K-DBQ variables did not show significant direct correlations with takeover time, their inclusion marginally improved the RF model's prediction accuracy. This suggests that K-DBQ responses capture latent behavioral tendencies that are not directly measurable. In ensemble learning, even weak predictors can enhance model stability and decision diversity when combined with stronger features. Therefore, while traditional self-report measures like the K-DBQ may not directly explain takeover performance, they still serve as valuable complementary predictors within data-driven frameworks. Incorporating driving experience and selected K-DBQ scores improved the RF classifier's performance by 13% compared to the baseline model using only the three statistically significant factors.

https://doi.org/10.3390/electronics15071390

We also found that prediction performance was improved by feature selection using RFE. The proposed RF classifier using only 10 key human factors achieved a 10% higher accuracy than the model using all 21 features. This demonstrates the importance of selecting optimal input variables before model development.

The proposed RF classifier can be incorporated into an advanced driver-assistance system (ADAS) that predicts slow-responding drivers and helps them reduce their takeover delay. We propose that the ADAS should include (1) a driver monitoring system that observes driver behavior, (2) an in-vehicle agent that provides timely and relevant situational information, and (3) the proposed RF classifier for driver takeover prediction. The system could predict low-performing drivers in advance and provide multimodal cues prompting them to stop NDRTs and prepare for upcoming TORs. Furthermore, the ADAS could monitor eye movements during automated driving and alert drivers to refocus on traffic or the in-vehicle agent if they are predicted to perform poorly. Older drivers with cognitive decline may benefit even more from such a system.

In addition, two important findings emerged. First, the K-DBQ—originally developed to measure manual driving behavior—can enhance takeover time prediction accuracy, highlighting that driver behavior questionnaires may still capture trait-level factors relevant to human–ADS interaction. Future studies should consider developing automation-specific behavioral questionnaires to better link habitual and situational factors. Second, the gender stereotype regarding female drivers' ability to regain vehicle control is unfounded and unsupported by our data.

**Author Contributions:** Conceptualization, J.K.; Methodology, J.K.; Software, J.K.; Validation, J.K.; Formal Analysis, J.K.; Investigation, J.K.; Data Curation, J.K.; Writing—Original Draft Preparation, J.K.; Writing—Review and Editing, J.K. and O.J.; Visualization, J.K.; Supervision, O.J. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by a National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. 2021R1A2C2095289), and, in part, by the Chungbuk National University BK21 program (2025).

**Data Availability Statement:** The data presented in this paper are available upon request from the corresponding author.

**Acknowledgments:** The dataset used in this study was originally collected under a project supported by a grant (20TLRP-B127450-04) from the Transportation and Logistics R&D Program funded by the Ministry of Land, Infrastructure and Transport of the Korean government (PI: Daesup Yoon).

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Sekadakis, M.; Yannis, G. Systematic review and meta-analysis of take-over time from automated driving at SAE levels 2 and 3 to manual control. *Transp. Res. Part F Traffic Psychol. Behav.* **2025**, *113*, 263–306. [CrossRef]
2. Zhou, F.; Yang, X.J.; Zhang, X. Takeover transition in autonomous vehicles: A YouTube study. *Int. J. Hum.–Comput. Interact.* **2020**, *36*, 295–306. [CrossRef]
3. *SAE J3016*; Standard—Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles. SAE International: Warrendale, PA, USA, 2021. Available online: https://www.sae.org/standards/j3016_202104-taxonomy-definitions-terms-related-driving-automation-systems-road-motor-vehicles (accessed on 13 March 2026).
4. Marberger, C.; Mielenz, H.; Naujoks, F.; Radlmayr, J.; Bengler, K.; Wandtner, B. Understanding and applying the concept of "driver availability" in automated driving. In *International Conference on Applied Human Factors and Ergonomics*; Springer International Publishing: Cham, Switzerland, 2017; pp. 595–605.
5. UN Regulation No. 157—Automated Lane Keeping Systems (ALKS). Available online: https://unece.org/transport/documents/2021/03/standards/un-regulation-no-157-automated-lane-keeping-systems-alks (accessed on 20 February 2026).
6. National Traffic safety and Environment Laboratory, Japan. Results of the Study on ACSF Transition Time. In Proceedings of the 4th Meeting of ACSF Informal Group, Tokyo, Japan, 25–27 November 2015.

7. Eriksson, A.; Stanton, N.A. Takeover time in highly automated vehicles: Noncritical transitions to and from manual control. *Hum. Factors* **2017**, *59*, 689–705. [CrossRef]
8. Wang, C.; Xu, C.; Shao, Y.; Zheng, N.; Peng, C.; Tong, H.; Xu, Z. Identifying factors affecting driver takeover time and crash risk during the automated driving takeover process. *J. Transp. Saf. Secur.* **2025**, *17*, 782–807. [CrossRef]
9. Wan, J.; Wu, C. The effects of lead time of take-over request and nondriving tasks on taking-over control of automated vehicles. *IEEE Trans. Hum.-Mach. Syst.* **2018**, *48*, 582–591. [CrossRef]
10. Wandtner, B.; Schömig, N.; Schmidt, G. Effects of non-driving related task modalities on takeover performance in highly automated driving. *Hum. Factors* **2018**, *60*, 870–881. [CrossRef]
11. Horáková, M.; Vondráčková, L.; Krejčí, L.; Trepáčová, M. Effects of non-driving-related tasks on driving performance after takeover request in automated driving (Level 3): A simulator study. *Eur. Transp. Res. Rev.* **2025**, *17*, 49. [CrossRef]
12. Hu, W.; Zhang, T.; Zhang, Y.; Chan, A.H.S. Non-driving-related tasks and drivers' takeover time: A meta-analysis. *Transp. Res. Part F Traffic Psychol. Behav.* **2024**, *103*, 623–637. [CrossRef]
13. Kim, H.; Kim, W.; Kim, J.; Lee, S.J.; Yoon, D. An analysis of transition characteristics of driver control authority according to NDRT type. In Proceedings of the International Symposium on Future Active Safety Technology toward Zero Accidents (FAST-Zero) 2019, Blacksburg, VA, USA, 9–11 September 2019; pp. 1–6.
14. Clark, H.; McLaughlin, A.C.; Feng, J. Situational awareness and time to takeover: Exploring an alternative method to measure engagement with high-level automation. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Austin, TX, USA, 9–13 October 2017*; SAGE Publications: Los Angeles, CA, USA, 2017; pp. 1452–1456.
15. Endsley, M.R. Measurement of situation awareness in dynamic systems. *Hum. Factors* **1995**, *37*, 65–84. [CrossRef]
16. Gold, C.; Körber, M.; Lechner, D.; Bengler, K. Taking over control from highly automated vehicles in complex traffic situations: The role of traffic density. *Hum. Factors* **2016**, *58*, 642–652. [CrossRef]
17. Du, N.; Yang, X.J.; Zhou, F. Psychophysiological responses to takeover requests in conditionally automated driving. *Accid. Anal. Prev.* **2020**, *148*, 105804. [CrossRef] [PubMed]
18. Li, S.; Blythe, P.; Guo, W.; Namdeo, A. Investigation of older driver's takeover performance in highly automated vehicles in adverse weather conditions. *IET Intell. Transp. Syst.* **2018**, *12*, 1157–1165. [CrossRef]
19. Kim, H.; Kim, W.; Kim, J.; Lee, S.-J.; Yoon, D.; Jo, J. A study on re-engagement and stabilization time on take-over transition in a highly automated driving system. *Electronics* **2021**, *10*, 344. [CrossRef]
20. Du, N.; Zhou, F.; Pulver, E.M.; Tilbury, D.M.; Robert, L.P.; Pradhan, A.K.; Yang, X.J. Predicting driver takeover performance in conditionally automated driving. *Accid. Anal. Prev.* **2020**, *148*, 105748. [CrossRef]
21. Liang, K.; Calvert, S.C.; Nordhoff, S.; Li, M.; van Lint, J. Predicting drivers' takeover time for safe and comfortable vehicle control transitions: The role of spare capacity and driver characteristics. *Appl. Ergon.* **2025**, *129*, 104603. [CrossRef] [PubMed]
22. Chen, H.; Zhao, X.; Li, H.; Gong, J.; Fu, Q. Predicting driver's takeover time based on individual characteristics, external environment, and situation awareness. *Accid. Anal. Prev.* **2024**, *203*, 107601. [CrossRef]
23. Espinoza-Molina, F.E.; Ortega, M.; Escobar, K.E.S.; Salazar, J.S.V. An integrated approach to the Spanish Driving Behavior Questionnaire (SDBQ) in the city of Cuenca, Ecuador. *Sustainability* **2024**, *16*, 4885. [CrossRef]
24. Alavi, S.S.; Mohammadi, M.; Soori, H.; Kalhori, S.M.; Sepasi, N.; Khodakarami, R.; Farshchi, M.; Hasibi, N.; Rostami, S.; Razi, H.; et al. Iranian Version of Manchester Driving Behavior Questionnaire (MDBQ): Psychometric Properties. *Iran. J. Psychiatry* **2016**, *11*, 37.
25. Useche, S.A.; Cendales, B.; Lijarcio, I.; Llamazares, F.J. Validation of the F-DBQ: A short (and accurate) risky driving behavior questionnaire for long-haul professional drivers. *Transp. Res. Part F Traffic Psychol. Behav.* **2021**, *82*, 190–201. [CrossRef]
26. Wickens, C.D. Processing resources and attention. In *Multiple Task Performance*; CRC Press: Boca Raton, FL, USA, 2020; pp. 3–34.
27. Kim, J.; Kim, W.; Kim, H.-S.; Yoon, D. A study on the correlation between subjective driver readiness and NDRT Type during automated driving. In *Proceedings of the 2020 International Conference on Information and Communication Technology Convergence (ICTC), Jeju, Republic of Korea, 21–23 October 2020*; IEEE: New York, NY, USA, 2020; pp. 1774–1776.
28. Dubes, R.C.; Jain, A.K. *Algorithms for Clustering Data*; Prentice Hall: Hoboken, NJ, USA, 1988.
29. Kim, J.; Kim, W.; Kim, H.-S.; Lee, S.-J.; Kwon, O.-C.; Yoon, D. A novel study on subjective driver readiness in terms of non-driving related tasks and take-over performance. *ICT Express* **2022**, *8*, 91–96. [CrossRef]
30. Reason, J.; Manstead, A.; Stradling, S.; Baxter, J.; Campbell, K. Errors and violations on the roads: A real distinction? *Ergonomics* **1990**, *33*, 1315–1332. [CrossRef] [PubMed]
31. Oh, J.S.; Lee, S.C. The structure of driving behavior determinants and its relationship between reckless driving behavior. *Korean J. Cult. Soc. Issue* **2011**, *17*, 175–197.
32. Smart Eye Pro 8.0. Available online: https://www.smarteye.se/ (accessed on 1 November 2025).
33. Rubio, S.; Diaz, E.; Martin, J.; Puente, J.M. Evaluation of subjective mental workload: A comparison of SWAT, NASA-TLX, and workload profile methods. *Appl. Psychol.* **2004**, *53*, 61–86. [CrossRef]

34. Hoaglin, D.C.; Iglewicz, B.; Tukey, J.W. Performance of some resistant rules for outlier labeling. *J. Am. Stat. Assoc.* **1986**, *81*, 991–999. [CrossRef]
35. Hoaglin, D.C.; Iglewicz, B. Fine-tuning some resistant rules for outlier labeling. *J. Am. Stat. Assoc.* **1987**, *82*, 1147–1149. [CrossRef]
36. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
37. Adelabu, S.; Mutanga, O.; Adam, E. Evaluating the impact of red-edge band from Rapideye image for classifying insect defoliation levels. *ISPRS J. Photogramm. Remote Sens.* **2014**, *95*, 34–41. [CrossRef]
38. Oumar, Z. Assessing the utility of the spot 6 sensor in detecting and mapping Lantana camara for a community clearing project in KwaZulu-Natal, South Africa. *South Afr. J. Geomat.* **2016**, *5*, 214–226. [CrossRef]
39. Abdel-Rahman, E.M.; Mutanga, O.; Adam, E.; Ismail, R. Detecting Sirex noctilio grey-attacked and lightning-struck pine trees using airborne hyperspectral data, random forest and support vector machines classifiers. *ISPRS J. Photogramm. Remote Sens.* **2014**, *88*, 48–59. [CrossRef]
40. Chan, J.C.-W.; Paelinckx, D. Evaluation of Random Forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. *Remote Sens. Environ.* **2008**, *112*, 2999–3011. [CrossRef]
41. Odindi, J.; Adam, E.; Ngubane, Z.; Mutanga, O.; Slotow, R. Comparison between WorldView-2 and SPOT-5 images in mapping the bracken fern using the random forest algorithm. *J. Appl. Remote Sens.* **2014**, *8*, 083527. [CrossRef]
42. Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene selection for cancer classification using support vector machines. *Mach. Learn.* **2002**, *46*, 389–422. [CrossRef]
43. Le, T.H.M.; Ali Babar, M. Mitigating data imbalance for software vulnerability assessment: Does data augmentation help? In Proceedings of the 18th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, Barcelona, Spain, 24–25 October 2024; pp. 119–130.
44. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
45. Liu, A. Acute Stress and Takeover Performance in Conditionally Automated Driving: Model Development and Validation Across Urban Road and Highway Contexts. Master's Thesis, University of Nottingham, Nottingham, UK, September 2025.
46. Saeb, S.; Lonini, L.; Jayaraman, A.; Mohr, D.C.; Kording, K.P. The need to approximate the use-case in clinical machine learning. *Gigascience* **2017**, *6*, gix019. [CrossRef] [PubMed]
47. Song, Y.; Wang, Z.; Wang, H.; Sun, G.; Gong, B.; Zhang, F. Interpretable deep learning for personalized energy expenditure prediction using ECG and acceleration signals in incremental exercise. *Sci. Rep.* **2025**, *15*, 36277. [CrossRef]
48. Oshiro, T.M.; Perez, P.S.; Baranauskas, J.A. How many trees in a random forest? In *Machine Learning and Data Mining in Pattern Recognition*; Perner, P., Ed.; Springer: Berlin/Heidelberg, Germany, 2012; pp. 154–168.
49. Kim, J.; Kim, M.; Park, K.; Kim, H.-S. Mental stress assessment using SVM with physiological sensor data. In Proceedings of the 2021 International Conference on Information and Communication Technology Convergence (ICTC), Jeju Island, Republic of Korea, 20–22 October 2021; IEEE: New York, NY, USA, 2021; pp. 1296–1299.
50. Matthews, B. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta (BBA)-Protein Struct.* **1975**, *405*, 442–451. [CrossRef]
51. Hanley, J.A.; McNeil, B.J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **1982**, *143*, 29–36. [CrossRef] [PubMed]
52. Skurichina, M.; Duin, R.P.W. Bagging, boosting and the random subspace method for linear classifiers. *Pattern Anal. Appl.* **2002**, *5*, 121–135. [CrossRef]
53. Fountas, G.; Pantangi, S.S.; Hulme, K.F.; Anastasopoulos, P.C. The effects of driver fatigue, gender, and distracted driving on perceived and observed aggressive driving behavior: A correlated grouped random parameters bivariate probit approach. *Anal. Methods Accid. Res.* **2019**, *22*, 100091. [CrossRef]
54. Monsell, S. Control of mental processes. In *Unsolved Mysteries of the Mind*; Psychology Press: Hove, UK, 2021; pp. 93–148.
55. Rogers, R.D.; Monsell, S. Costs of a predictible switch between simple cognitive tasks. *J. Exp. Psychol. Gen.* **1995**, *124*, 207. [CrossRef]
56. Jersild, A.T. *Mental Set and Shift*; Columbia University: New York, NY, USA, 1927.
57. Allport, A.; Styles, E.A.; Hsieh, S. 17 Shifting Intentional Set: Exploring the. In *Attention and Performance XV: Conscious and Nonconscious Information Processing*; MIT Press: Cambridge, MA, USA, 1994; Volume 15, p. 421.
58. Pashler, H.; Johnston, J.C.; Ruthruff, E. Attention and performance. *Annu. Rev. Psychol.* **2001**, *52*, 629–651. [CrossRef] [PubMed]
59. Anstey, K.J.; Wood, J. Chronological age and age-related cognitive deficits are associated with an increase in multiple types of driving errors in late life. *Neuropsychology* **2011**, *25*, 613. [CrossRef] [PubMed]