Accident Analysis and Prevention 183 (2023) 106956

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

On the forces of driver distraction: Explainable predictions for the visual demand of in-vehicle touchscreen interactions

Patrick Ebel [a,*], Christoph Lingenfelder [b], Andreas Vogelsang [a]

[a] University of Cologne, Cologne, Germany
[b] MBition GmbH, Berlin, Germany

ARTICLE INFO

Keywords:
Driver distraction
Visual demand
In-vehicle information systems
Naturalistic driving data
Touchscreen interactions

ABSTRACT

With modern infotainment systems, drivers are increasingly tempted to engage in secondary tasks while driving. Since distracted driving is already one of the main causes of fatal accidents, in-vehicle touchscreens must be as little distracting as possible. To ensure that these systems are safe to use, they undergo elaborate and expensive empirical testing, requiring fully functional prototypes. Thus, early-stage methods informing designers about the implication their design may have on driver distraction are of great value. This paper presents a machine learning method that, based on anticipated usage scenarios, predicts the visual demand of in-vehicle touchscreen interactions and provides local and global explanations of the factors influencing drivers' visual attention allocation. The approach is based on large-scale natural driving data continuously collected from production line vehicles and employs the SHapley Additive exPlanation (SHAP) method to provide explanations leveraging informed design decisions. Our approach is more accurate than related work and identifies interactions during which long glances occur with 68 % accuracy and predicts the total glance duration with a mean error of 2.4 s. Our explanations replicate the results of various recent studies and provide fast and easily accessible insights into the effect of UI elements, driving automation, and vehicle speed on driver distraction. The system can not only help designers to evaluate current designs but also help them to better anticipate and understand the implications their design decisions might have on future designs.

1. Introduction

Nowadays, large center stack touchscreens, like the ones found in Tesla's Model 3[1] or the Mercedes-Benz EQS[2] are the main interface between the driver and the In-Vehicle Information Systems (IVISs). During the interaction, drivers need to take their eyes off the road to scan the information presented on the screen. Thus, they distribute their visual attention between the primary driving task and the secondary touchscreen task. This increases the risk of a crash significantly (Dingus et al., 2016; Green, 1999), in particular for eyes-off-road durations longer than two seconds (Klauer et al., 2006). With IVISs becoming more complex and incorporating an ever-increasing amount of functionalities, drivers have more options than ever to interact with them while driving. The temptation to engage in non-driving related tasks is further increased by constantly improving driving automation features. During partially automated driving, drivers tend to engage more often in non-driving related tasks, even though they are still supposed to monitor the vehicle (Carsten et al., 2012; de Winter et al., 2014; Ebel et al., 2022). To ensure that IVISs are safe to use, they are subject to strict regulations and elaborate test protocols. Automotive Original Equipment Manufacturers (OEMs) conduct expensive empirical user studies in artificial settings (e.g., driving simulators) to test the safety of the systems. However, driving simulator studies can only replicate real-world driving behavior to a certain degree (Riener, 2011) and often lack absolute validity (Kaptein et al., 1996; Fisher, 2011). Furthermore, to evaluate a new IVISs design in a user study, all relevant features need to be implemented in a functional prototype. Although such measures will remain necessary, automotive UX experts require explainable evaluation methods (Ebel et al., 2020) allowing them to identify potentially distracting interaction patterns already in the early design stages. Such automated methods can facilitate the development of interaction concepts that are safe by design and, therefore, less likely to fail final evaluations.

* Corresponding author.
E-mail addresses: ebel@cs.uni-koeln.de (P. Ebel), christoph.lingenfelder@mercedes-benz.com (C. Lingenfelder), vogelsang@cs.uni-koeln.de (A. Vogelsang).

[1] https://www.tesla.com/model3

[2] https://www.mercedes-benz.com/en/innovation/future-mobility/eqs-with-unique-mbux-hyperscreen/

https://doi.org/10.1016/j.aap.2023.106956
Received 28 July 2022; Received in revised form 3 January 2023; Accepted 5 January 2023
Available online 19 January 2023

Current approaches that predict driver distraction based on user interaction information are derived from methods like Fitt's Law (Fitts, 1954) or the Keystroke-Level Model (KLM) (Card et al., 1980), where the time of certain operations is summed up to predict the total time of a task. The glance behavior is then derived from this metric. Although these models are limited by their cumulative linear nature and require experts to manually specify each task, they are highly interpretable. With more complex models, interpretability is often sacrificed for increased accuracy, highlighting the inherent trade-off between the two (Lundberg and Lee, 2017). However, without explanations, scientific findings may remain hidden, user acceptance suffers, and the learning effect is limited (Molnar, 2020; Doshi-Velez and Kim, 2017). To facilitate data-informed design decisions, it is not only important to predict potentially dangerous interaction sequences but also to understand which interactions force drivers to take their eyes off the road.

2. Background and related work

In this section, we introduce the concept of visual demand. We explain how to measure it and present computational models of visual demand. Finally, we introduce SHAP, an approach to generate explanations for machine learning predictions.

2.1. Visual demand of secondary task engagements

Visual demand is the *"degree or quantity of visual activity required to extract information from an object to perform a specific task"* (ISO15007, 2020). While driving a car, visual distraction from the primary driving task by engaging in a secondary task, compromises driving performance and safety (Engström et al., 2005; Donmez et al., 2010; Liang and Lee, 2010; Green, 1999; Klauer et al., 2006; Horrey and Wickens, 2007). This also applies to higher automation levels of driving automation (SAEJ3016, 2021). Recent research shows that takeover performance after a stretch of automated driving is significantly affected by the visual-cognitive load of the secondary task (Wintersberger et al., 2021) and distraction in general (Merlhiot and Bueno, 2021). In ISO:15007:2020 (ISO15007, 2020) multiple metrics to measure visual demand are described. Two of the metrics that are widely used are the Total Glance Duration (TGD) and the average glance duration. The TGD is the "summation of all glance durations to an area of interest (or set of related areas of interest) during a condition task, subtask or sub-subtask". The average glance duration is the "mean duration of all glance durations to an area of interest (or set of related areas of interest) during a condition task, subtask or sub-subtask)". Further research (Horrey and Wickens, 2007) shows that single longer-than-normal glances, especially those longer than two seconds (Klauer et al., 2006), highly correlate with reduced driving safety. This is also mentioned in the "Visual-Manual NHTSA Driver Distraction Guidelines for In-Vehicle Electronic Devices" (NHTSA, 2012). However, according to Victor et al. (2014) there is not a single metric that can fully describe the relationship between glance behavior and risk, but rather a combination of metrics is necessary. Burns et al. (2010) additionally argue that whereas any single measure only provides an incomplete assessment of distraction, empirically supported measures such as the above-introduced ones should be included for decision-making as early as possible in the design process.

2.2. Visual demand prediction

Various methods aim to predict visual-manual distraction while driving (Kanaan et al., 2019; Li et al., 2018; Wollmer et al., 2011; Li et al., 2020; Risteska et al., 2021). Most of them focus on driver distraction detection to warn the driver when a potentially dangerous situation is detected. These approaches are often based on naturalistic driving data and employ various machine learning methods. They utilize driving performance metrics (e.g., speed or steering wheel angle) (Kanaan et al., 2019; Li et al., 2018; Wollmer et al., 2011), environmental data (e.g., traffic conditions) (Risteska et al., 2021), or video data of the driver (Li et al., 2020; Kutila et al., 2007). While these approaches show promising results, they do not incorporate any information on how drivers interacted with secondary devices like mobile phones or IVISs. Therefore, they do not generate insights into the visual demand of specific UI elements or interactions.

However, various approaches exist that model the visual demand of in-vehicle Human-Machine Interfaces (HMIs) based on user interactions with specific UI elements. They explain the effect specific interactions have on drivers' visual distraction. These approaches focus on the understanding of interaction behavior and aim to identify distracting features of in-vehicle HMIs. Their purpose is to inform designers and researchers in the early stages of the development process about possible implications their design might have on driver distraction. In this work, we focus on the latter and provide an overview of the current state-of-the-art in this domain.

Most of the approaches that predict visual demand, based on user interactions, are bottom-up approaches derived from the KLM modeling technique (Card et al., 1980; Card, 1983). In such approaches, an entire task is decomposed into a sequence of specific primitive operators (e.g., pressing a button, or searching in a list). The interaction durations for each operator are then determined empirically (Schneegaß et al., 2011). The overall time on task predictions are then equal to the sum of the individual interaction durations of the respective operators occurring in the task. The KLM technique was originally developed to predict processing times in computer-assisted office work, but multiple adjustments were made to assess IVISs (Schneegaß et al., 2011; Manes, 1997; Lee et al., 2019). However, most of these approaches focus on task completion times rather than visual demand. Pettitt et al. (2007) were the first to propose a KLM-based approach to predict visual demand. They show a high correlation between predicted values and measures from an occlusion experiment. The first KLM-based method to directly predict visual demand is presented by Purucker et al. (2017), who propose a task-specific KLM model. They argue that using fixed operators to model innovative and new hardware can only work to a limited extent. Whereas their approach can only predict TGD, Large et al. (2017b) propose a method that can additionally predict the number of glances and the mean glance duration. Their information-theoretic approach is based on the Hick–Hyman Law for decision/search time and Fitt's Law for pointing time. Whereas the results achieved by the presented KLM-based approaches are promising, they all share several drawbacks. First, due to their cumulative and linear character, the models are not suited to model potential (non-linear) dependencies between different user interactions or driving situations. For example, the difference in the visual demand between selecting an element out of a list and tapping a button might be negligible for lower speeds but significant for higher speeds. Additionally, the length of an interaction sequence in combination with specific interactions might also influence visual demand in a non-linear and non-additive way (Purucker et al., 2017). For example, if the driver presses two buttons that are located close to each other, it unlikely results in a doubling of the TGD as the driver might perform both interactions during one glance. Second, the model parameters of the introduced approaches are derived from empirical testing in restricted driving scenarios using driving simulators of different fidelity, and a relatively small number of participants. This can likely lead to predictions being very context-dependent, as also noted by Large et al. (2017b) and shown in a real-world driving experiment evaluating the applicability of Fitt's Law (Pampel et al., 2019). Third, current approaches do not consider the effect the driving situation has on the visual demand. Research shows that drivers modulate their task engagement and visual attention based on driving demands (Risteska et al., 2021) and the degree of assisted driving (Morando et al., 2021; Tsimhoni and Green, 2001; Gaspar and Carney, 2019; Large et al., 2017a) making it important to include such parameters.

A different approach is taken by Kujala and Salvucci (2015) who propose a model based on the ACT-R cognitive model architecture (Anderson et al., 2004). Their approach aims to represent the visual sampling strategy of drivers. They argue that drivers adjust their glances based on a time limit that is dependent on the current driving performance. Whereas the model can predict multiple facets of visual demand, only grid and list layouts are considered. Furthermore, the driving scenario is fairly simple and the evaluation shows significant drawbacks in prediction accuracy, especially concerning the detection of long glances.

We argue that the utilization of large natural driving and interaction data in combination with machine learning approaches that can model non-linear relationships can be a promising step toward more accurate and holistically applicable solutions. Machine learning approaches have led to high-quality predictions in the domain of distraction detection methods as introduced above.

However, two main factors prevent these approaches from generating valuable insights into how specific design elements affect visual demand. First, interaction data is not yet available in a similar quantity as driving data. Second, all the above-presented machine learning approaches lack explainability. Whereas certain performance metrics are reported, the models remain a black box without providing insights on the features that are decisive for the predictions.

2.3. Explainable predictions with SHAP

Explainable AI (XAI) aims to make machine learning models more transparent by providing human-understandable (interpretable) information, explaining the behavior and processes of machine learning models (Barredo Arrieta et al., 2020; Liao et al., 2020). Explanations serve as an interface between the human and the model (Barredo Arrieta et al., 2020) and can be valuable in various applications (Wiegand et al., 2020; Wang and An, 2021). Explanations can enhance scientific understanding (Doshi-Velez and Kim, 2017), increase user trust (Shin, 2021; Lipton, 2018), and can help to infer causal relations in data (Verma et al., 2020). For the task at hand explainable predictions are of particular interest because the goal is not only to make predictions of the visual demand but also to draw conclusions about the impact of specific UI elements, gestures, and varying driving situations. The goal of this approach is to enable AI-assisted decision-making (Zhang et al., 2020), optimizing a joint decision based on the domain knowledge of the human expert and the insights generated by the model prediction and accompanying explanation.

SHAP, proposed by Lundberg and Lee (2017) is a method based on Shapley values from coalitional game theory (Shapley, 1953). The SHAP method provides local and global explanations for arbitrary predictive models. SHAP belongs to the class of additive feature attribution methods. The main idea is to use an interpretable explanation model $g(z')$ in the form of a linear function such that the model's prediction of a certain instance is equal to the sum of its feature contributions $\phi_i \in \mathbb{R}$ (Molnar, 2020):

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z_i', \tag{1}$$

where $z' \in \{0,1\}^M$ with z_i' represents the presence of feature i, ϕ_0 represents the models output in case no feature is present, and M is the number of input features (Lundberg et al., 2020).

Lundberg and Lee (2017) further state that a single unique solution exists that follows the definition of additive feature attribution methods (see (1)) and satisfies the properties of local accuracy, missingness, and consistency. Local accuracy describes that the sum of the feature attributions is equal to the prediction of the original model. Missingness describes that a missing feature ($z_i = 0$) gets assigned an attribution of zero and consistency states that when changing a model such that it is more dependent on a certain feature, the attribution of that feature should not decrease.

The only possible solution as described by Lundberg and Lee (2017) is given by the SHAP values:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|M| - |S| - 1)!}{M!} \left(f_x(S \cup \{i\}) - f_x(S) \right), \tag{2}$$

with S being the set of non-zero indexes in z', $f_x(S)$ being the expected value of the function conditioned on a subset S of the input features, and N being the set of all input features.

Multiple different approaches exist to approximate SHAP values for different kinds of machine learning models. However, in this study, we use TreeSHAP (Lundberg et al., 2020) which allows the computation of exact SHAP values for tree-based approaches.

Compared to approaches like LIME (Ribeiro et al., 2016) or approaches specific to tree-based models like permutation importance of feature impurity calculations, SHAP has many advantages. Due to the solid foundation in game theory (Molnar, 2020), SHAP values come with theoretical guarantees about consistency and local accuracy. Additionally, local and global explanations are consistent, SHAP values indicate whether the contribution of each feature is positive or negative, and Lundberg et al. (2018) show a greater overlap of SHAP values and human intuition (Lundberg and Lee, 2017; Lundberg et al., 2020).

3. Proposed approach

In this work, we showcase how to effectively use machine learning methods to predict and explain the visual demand of in-vehicle touchscreen interactions based on large naturalistic driving data. The contribution of this paper is two-fold: First, we propose a machine learning approach predicting the visual demand of in-vehicle touchscreen interactions based on the type of interaction and the associated driving parameters. Second, we apply the SHAP method (Lundberg and Lee, 2017) to explain the predictions and to visualize how user interactions, vehicle speed, steering wheel angle, and automation level, affect drivers' long glance probability and TGD. In the following, we introduce several definitions that will be used in the remainder of the paper. Furthermore, we describe the data collection, data processing, and modeling procedures in detail.

3.1. Definitions and problem statement

The goal of this approach is to predict drivers' visual attention allocation based on user interactions and the associated driving parameters. To do so, we model drivers' secondary task engagements by combining interaction sequences, driving sequences, and glance sequences. These concepts are introduced in the following.[3]

Interaction Sequence. An interaction sequence is defined as a set of subsequent touchscreen interactions performed by the driver. The duration between two subsequent interactions must be smaller than Δt_{max}.

Glance Sequence. A glance sequence is defined as a set of subsequent driver glances toward a predefined Area of Interest (AOI).

Driving Sequence. A driving sequence is a sequence of driving data observations. Each observation consists of the vehicle speed, the steering wheel angle, and the status of Adaptive Cruise Control (ACC) and Steering Assist (SA).

Secondary Task Engagement. A secondary task engagement S describes the touchscreen interactions, the driving behavior, and the glance behavior of a driver while interacting with the center stack touchscreen. We consider the vehicle speed and steering wheel angle from t_b seconds before the first until t_b seconds after the last interaction. Furthermore, all glances that fall in between the first and last interaction are considered.

[3] For the formal definitions refer A.1.

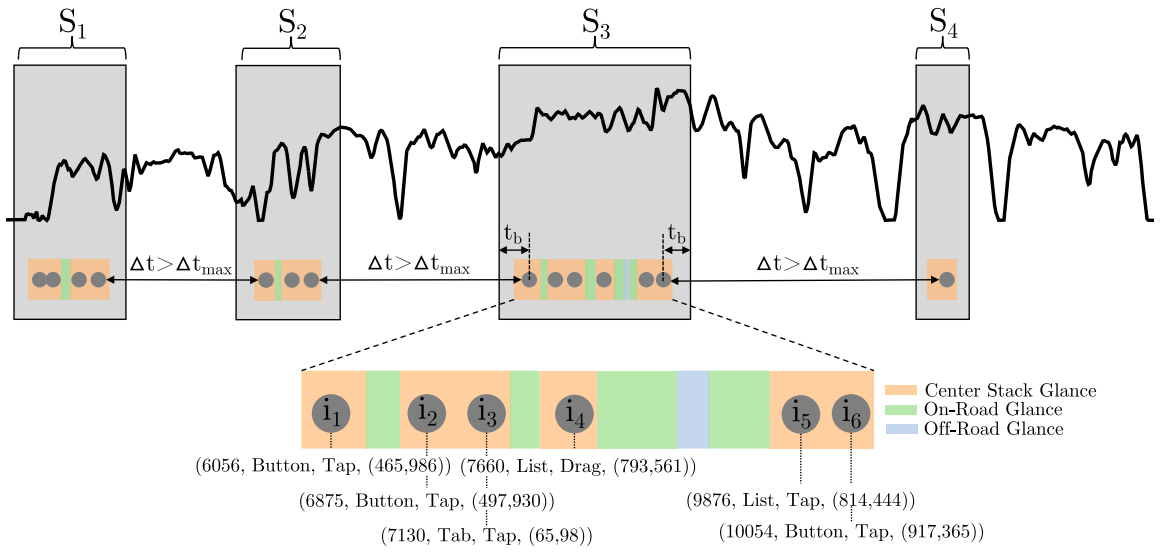


Fig. 1. Schematic overview on how secondary task engagements S_n are extracted from driving sequences (solid black line), glance sequences (colored rectangles), and interaction sequences (gray dots). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Long Glance Prediction Task. The long glance prediction task describes the task of predicting if the driver will look at the center stack touchscreen for more than two seconds during a secondary task engagement.

TGD Prediction Task. The TGD prediction task describes the task of predicting the drivers' total glance duration toward the center stack touchscreen during a secondary task engagement.

3.2. Data collection

The data used in this study was collected over the air from over 100 test cars and five different car models via the Mercedes-Benz telematics data logging framework. A more detailed description of the logging mechanism is provided by Ebel et al. (2021a). The data collection period ranged from mid-October 2021 to mid-January 2022. All company internal test cars with the latest software architecture, an eye-tracker, and ACC and SA functionality, contributed to the data collection. The test cars were used for various, mostly not UI-related test drives, but also for leisure drives of employees. While the cars were driven all over Europe, most of the trips were recorded in Germany. We leveraged the data collection and processing framework of Mercedes-Benz, to collect touchscreen interactions, driving data (vehicle speed, steering wheel angle, and level of driving automation), and eye-tracking data. We did not collect any demographic or environmental data.

The touchscreen interactions were collected from the UI software, where a datapoint was logged each time the driver touched the center stack touchscreen. A datapoint consists of the touched UI element, the start and end position of all fingers used for the gesture, and a timestamp.

The glance data was acquired via a stereo camera located in the instrument cluster behind the steering wheel. The system is already commercialized and available in production line vehicles. As with most remote eye trackers, gaze detection is based on the pupil-corneal reflection technique (Merchant, 1967). In this method, the pupil center is tracked in relation to the position of the corneal reflection (Hutchinson et al., 1989). The driver's field of view is divided into different AOIs and a new glance is collected each time the focus of the driver switches between AOIs. Across all conditions, the average true positive rate for each of the AOIs is above 90 percent. The system captures no raw video data.

All driving-related data was directly collected from the Controller Area Network (CAN) bus. The continuous signals (vehicle speed and steering wheel angle) were collected with a frequency 4 Hz, and the event-based signals (ACC, SA, and seat belt information) were collected on change. ACC automatically adjusts the vehicle speed based on speed limits and the vehicles ahead. SA actively supports lateral control to keep the car centered in the lane. Both systems operate at speeds between 0 km/h and 210 km/h.

3.3. Data processing

In the following we describe how we process the data such that the respective sequences follow the definitions given in 3.1. Fig. 1 shows a schematic overview describing the data synthesis of the individual sequences. For clarity, we only display the vehicle speed (solid black line), representative of the other driving data.

3.3.1. Interaction data

In controlled experiments, participants are instructed to perform pre-defined tasks that are specified by the experimenter. In such settings, it is straightforward to map user interactions and tasks. However, in the observational setting at hand, no task boundaries are defined. We do not know which task the driver intended to complete during the secondary task engagement, nor are the start or end times defined. One possibility to extract interaction sequences would be to consider all interactions that occurred during one trip. However, this would lead to very long interaction sequences with dense clusters of interactions sparsely distributed over a long period of time. To solve this problem, we set the maximum update interval, as defined in 3.1 to 10 s, $\Delta_{t_{max}} = 10$ s. We assume that after a period of 10 s with no interaction the driver disengaged from the secondary task and the interaction sequence ended after the last interaction. We then consider the next interaction as the starting point of a new interaction sequence. We argue that the 10-second assumption is valid, because both the distribution of interaction sequence durations and the distribution of total glance times toward the center stack touchscreen match well with values reported in the literature (NHTSA, 2012; Angell et al., 2008).

Table 1
Overview of the final input features describing a secondary task engagement.

Feature	Description
Interaction Data	
n_{Button}	# Interactions with regular buttons (e.g., push or radio buttons)
n_{List}	# Interactions with lists (e.g., when choosing a suggested destination)
n_{Map}	# Interactions with a map viewer (e.g., when zooming or dragging the navigation map)
n_{Slider}	# Interactions with slider elements (e.g., when changing the volume)
n_{Homebar}	# Interactions with the static homebar on the bottom of the screen
$n_{\text{CoverFlow}}$	# Interactions with cover flow widgets (e.g., when scrolling through albums covers)
n_{AppIcon}	# Interactions with app icons on the home screen
n_{Tab}	# Interactions with tab bars
n_{Keyboard}	# Interactions on the keyboard or number pad (e.g., when entering a destination)
n_{Browser}	# Interactions within the web browser
n_{RemoteUI}	# Interactions within Apple Car Play or Android Auto
$n_{\text{ControlBar}}$	# Interactions with a control bar, displayed as a small overlay on various screens
n_{PopUp}	# Interactions with pop-up element
$n_{\text{ClickGuard}}$	# Interactions with non-interactable background elements
n_{Other}	# Interactions with a UI element that does not fit any of the above categories
n_{Unknown}	# Interactions with a UI element for which the identifier could not be extracted
n_{Tap}	# Tap gestures
n_{Drag}	# Drag gestures
$n_{\text{Multitouch}}$	# Multitouch gestures
d_{avg}	Average distance between two consecutive interactions in *px*
N	Number of interactions
Driving Data	
v_{avg}	Average vehicle speed in km/h
θ_{avg}	Average steering wheel angle in °
a_{acc}	Status of the adaptive cruise control $a_{\text{acc}} \in \{0, 1\}$
a_{sa}	Status of the steering assist $a_{\text{sa}} \in \{0, 1\}$

For all remaining interactions, we compute the gesture type (*Tap, Drag, and Multitouch*) and distance between two interactions from the positioning information of the fingers. Finally, we map each UI element to an overarching element type (see Table 1). This step reduces data sparsity and ensures that the approach can produce generalizable statements about specific element types.

3.3.2. Glance data

We apply multiple filtering steps to improve the data quality of the glance data. The filtering is partially adapted from related research (Morando et al., 2019). In the first step, the glance information is aggregated into broader AOIs (*On-Road, Off-Road, Center Stack*). According to ISO 15007-1:2020 (ISO15007, 2020), we consider all glances that are not directly directed on the road (e.g., glances in the rear-view mirrors) as off-road glances. As we are explicitly interested in glances toward the center stack touchscreen, we distinguish these glances from regular off-road glances. Second, as described in Section 3.1, we consider all glances between the first and last touchscreen interaction of the associated interaction sequence. Fragmented glances at the beginning or end of an interaction sequence that start before the first interaction or end after the last interaction are considered as a whole. Third, short periods (less than 300 ms) of tracking loss are interpolated if the AOI preceding the loss is equal to the one succeeding it. Loss of tracking can occur due to changing lighting conditions, reflections in glasses, or when the camera view is blocked by the driver's hands on the steering wheel. Fourth, according to ISO 15007-1:2020 (ISO15007, 2020) glances shorter than 120 ms are interpolated because shorter fixations to an area of interest are physically not possible. Fifth, following the same argumentation, loss of tracking between different AOIs shorter than 120 ms is interpolated as well. Sixth, eyelid closures shorter than 500 ms are interpolated to remove blinks as suggested by ISO 15007-1:2020 (ISO15007, 2020).

3.3.3. Driving data

The driving data consists of continuous data and event-based data. In the first step, we extract the data that is relevant for the associated interaction sequences. To get a more stable estimate of the driving parameters in case of very short interaction sequences that might only consist of a single interaction, we consider steering wheel and vehicle speed data starting two seconds before the first interaction until two seconds after the last interaction of an interaction sequence ($t_b = 2$ s). After data extraction, sequences with missing values or error values, and sequences that show deviations in the logging frequency are discarded.

3.3.4. Data aggregation and final filtering

In total, we extracted 322,425 touchscreen sequences. We obtained valid speed data for 145,973 sequences, valid steering data for 81,150 sequences, and valid glance data for 111,792 sequences. After individual processing, we computed the intersection of the individual data sources resulting in 30,158 complete secondary task engagements. Most of the sequences were excluded either because they were generated on a test bench (no driving data was available), the car was not equipped with a camera, or because the sampling requirements were violated due to a loss of data connection. In the second stage of data processing, we apply further filtering steps to increase data quality. To prevent the data from becoming too sparse, we discarded 342 secondary task engagements with more than 41 interactions ($N > 41$ corresponds to the 99th percentile of the distribution of N). These secondary task engagements can be considered outliers without providing additional benefits for the use case at hand. We further discard 16,864 secondary task engagements where a passenger was present because they also tend to interact with the center stack touchscreen. These interactions cannot be mapped to driver glances and would skew the data toward fewer and shorter glances per secondary task engagement with many interactions logged that did not originate from the driver. Furthermore, we discarded 809 engagements during which the car came to a full stop and one sequence due to a remaining speed error. After this processing step, we obtained the final set of 12,142 secondary task engagements. Finally, we compute summary statistics for the secondary task engagements to generate the final set of features as described in Table 1. These features serve as input to the models introduced in the following.

3.4. Modeling

As formulated in the problem statement we solve one classification task (long glance prediction) and one regression task (TGD prediction). For each of the tasks, we compare a *Baseline* approach and a *Logisitc/Linear Regression* approach against three machine learning approaches, namely *Random Forests*, *Gradient Boosting Trees*, and *Feed-forward Neural Networks (FNNs)*. In the long glance prediction task, the *Baseline* approach randomly predicts one of the two classes (i.e. in a balanced dataset the probability of correctly predicting a long glance is roughly 50 %). In the TGD prediction task, the baseline approach predicts the median TGD of the training dataset. The parameters of the machine learning-based methods are chosen based on extensive hyperparameter optimization using random search.[4]

4. Evaluation

In this section we present the final dataset, put the experimental results in perspective, and elaborate on the explainable predictions generated by applying the SHAP method.

4.1. Dataset

The final dataset consists of 12,142 secondary task engagements sampled from 3,046 individual trips. The majority of secondary task engagements were collected from the Mercedes-Benz S-Class (7,342 secondary task engagements), EQS (3604), and EQE (824) models. The cars were equipped with a 17.7", 12.8", or 11.9" center stack touchscreen with similar pixel density. In total, 61,943 touch interactions and 119,770 individual glances were collected. The median trip length is 34.28 min ($Q_1 = 17.49, Q_3 = 66.58$). Specific glance and interaction statistics of the final dataset are presented in Fig. 2.[5]

In Figs. 2(e) and 2(f), the glance duration distribution during secondary task engagements (blue) is plotted against the glance duration distribution over all sessions independent of the driver being engaged in a secondary task (orange). This allows a comparison with approaches that utilize data collected irrespective if the driver being engaged in a secondary task or not.

We further compare our data with the manual driving baseline of the *100-Car Study* (Dingus et al., 2006) (data provided by Custer (2018), the *SHRP2* (Victor et al., 2014) (data available in Bärgman et al. (2015)) and the data reported in the work of Morando et al. (2019) (provided by the authors upon request). Figs. 2(h) and 2(i) show the glance distribution of on-road and off-road glances for the respective datasets. The glance distributions were truncated at 6 s since this corresponds to the length of the segments in the 100-car baseline dataset. The visual comparison shows that the off-road glance duration distribution matches well with the data reported in the three related studies. However, the on-road glances show some differences between our data and the data reported in the 100-Car study and the study of Morando et al. (2019). Whereas the mode is similar for all three datasets, the on-road glances in our study tend to be shorter compared to the other two studies. The potential reasons for this are manifold. For example, Morando et al. (2019) only consider driving segments of very controlled driving by excluding curved driving, lane changes, and driving segments with a vehicle speed under 60 km/h. In these rather calm driving situations, drivers need to switch less often between the road and off-road regions such as mirrors or side windows resulting in longer continuous on-road glances. The differences with regard to the 100-Car study could be due to the fact that the data is now almost 20 years old and only covers manual driving. The technology of the vehicles at that time, and in particular that of the infotainment and assistance systems, was fundamentally different from that in today's vehicles. However, considering the differences in the data collection, the comparison suggests that our data collection and processing pipeline produces representative data.

Fig. 2(e) indicates that during touchscreen] interactions, drivers need to distribute their visual attention between the road and the center stack resulting to shorter on-road glances. On the other hand, center stack glances during secondary task engagements tend to be longer than general center stack glances (Fig. 2(f)). Through Fig. 2(b), we see that roughly 25 % of all sequences consist of only a single interaction. This results in many short secondary task engagements that only consist of a single glance toward the center stack (Fig. 2(d)). These short engagements are part of real-world user behavior. However, they are often not represented in laboratory studies where only a few predefined tasks are evaluated. We argue that it is still relevant to analyze these short engagements and therefore decide to consider them. For the long glance classification task, we balanced the dataset by applying random undersampling. The resulting dataset consists of 4816 sequences for each class.

4.2. Experimental results

We evaluate the regression models using a repeated 10-fold cross-validation (Kohavi, 1995) and the classification models using a stratified 10-fold cross-validation. The results are given in Table 2. The models were fitted on the full set of input features given in Table 1.

The machine learning-based approaches outperform the Baseline approach and the Logistic and Linear Regression approaches in both tasks. The differences in the prediction accuracy support our assumption that neither of the problems at hand can be considered a linear problem and that interaction effects between different features exist. The machine learning models provide similar results. However, the Random Forest approaches offer two desirable properties making them in particular suitable for the use case at hand. First, the TreeSHAP (Lundberg et al., 2020) algorithm allows efficient computation of exact SHAP values for Random Forest models. Second, Random Forests can be run in parallel, making them suitable for future use cases when they are deployed on data of a whole production fleet. Thus, we choose the Random Forest models for the following explanation generation.

4.3. Explainable predictions

While the above-presented results provide a good measure of prediction accuracy, they are of limited value when it comes to understanding human behavior. To truly support researchers and practitioners in the design process to foster a deeper understanding of drivers' visual attention allocation, it needs more than just predicting whether a new user flow might cause too much distraction (Ebel et al., 2021b). For this reason, we employ SHAP. SHAP values represent the features' contribution to the model's output, providing a local explanation for each input sample. By combining many local explanations, one can represent global structures producing detailed insights into model behavior (Lundberg et al., 2020).

4.3.1. Local explanations

Fig. 3 displays the explanations for one long glance prediction and one TGD prediction. These force plots represent a particular model output as a cumulative effect of feature contributions (i.e. SHAP values). The length of each bar indicates how much the associated feature value pushes the model output from the base value toward higher values (red, to the right) or lower values (blue, to the left). The base value is computed as the average model output over the training dataset. The features in each group are sorted based on the magnitude of their impact and only the most influential features are displayed. The feature values are shown below the bars. For the long glance prediction, feature

[4] For more details on the search space refer to: A.2.
[5] For the full dataset statistics refer to A.3.

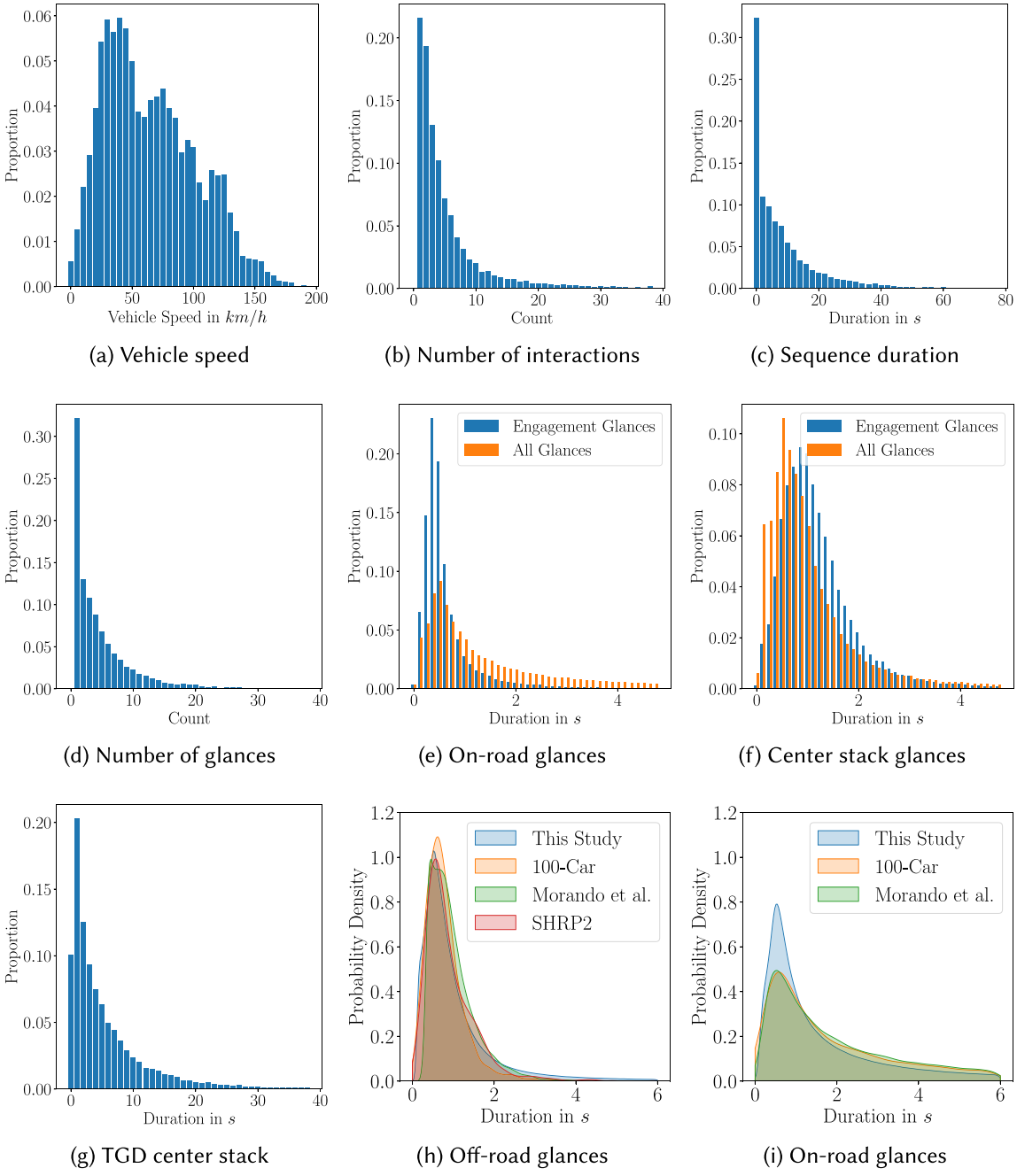
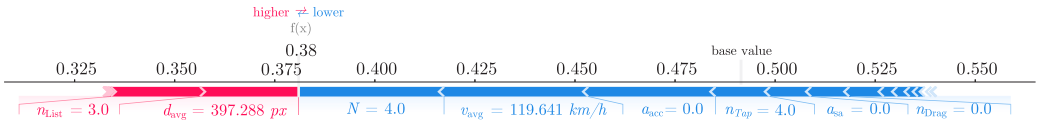


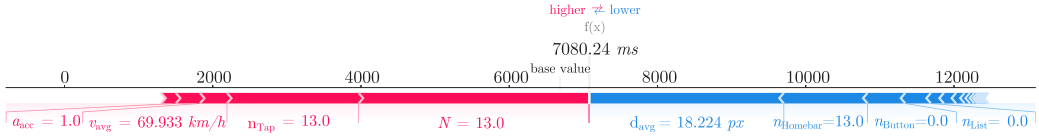
Fig. 2. Histograms that visualize the glance and interaction data. They show the (a) average speed per secondary task engagement, (b) the number of interactions per secondary task engagement, (c) the duration of the interaction sequences of a secondary task engagement, and the (d) number of glances toward the touchscreen during per secondary task engagement. Figure (e) compares the on-road glance duration distribution for glances during a secondary task engagement with glances irrespective whether a touchscreen interaction was performed or not. Figure (f) establishes the same comparison for glances toward the center stack touchscreen. Figure (g) shows the total glance duration toward the center stack touchscreen during a secondary task engagement. Figure (h) and (i) compare the probability density functions of the off-road and on-road glance duration from this study with the 100-Car study, the SHRP2 study (only off-road glances), and the study of Morando et al. (2019).

Table 2
Comparison of the different models.

Model	Long Glance Prediction		Total Glance Duration Prediction	
	Accuracy	Standard Deviation	Mean Absolute Error	Standard Deviation
Baseline	50.09 %	1.63 %	4378 ms	177 ms
Logistic/Linear Regression	61.93 %	1.69 %	3778 ms	383 ms
Random Forest	**67.53 %**	**1.38 %**	**2437 ms**	**112 ms**
XGBoost	67.22 %	1.85 %	2385 ms	117 ms
FNN	65.90 %	2.02 %	2443 ms	109 ms



(a) Long Glance Prediction Sample.



(b) Total Glance Duration Prediction Sample.

Fig. 3. Local explanations visualized as force plots.

contributions are displayed as probabilities. For the TGD prediction, they are shown in milliseconds.

Fig. 3(a) visualizes the explanation of a secondary task engagement for which the model outputs a long glance probability of 0.38 = 38 %. The long glance probability is pushed to the left because the driver only performed 4 interactions ($N = 4$) and drove at a speed of v_{avg} = 119.641 km/h while the ACC was deactivated ($a_{acc} = 0$). On the other hand, the prediction is pushed to the right because the interactions were quite distributed over the screen ($d_{avg} = 397.288$ px) and three of them were list interactions ($n_{List} = 3$).

Another secondary task engagement is explained in Fig. 3(b). Here, the TGD prediction of roughly 7 s is close to the base value because the positive and negative feature contributions balance each other out. During this secondary task engagement, the driver performed 13 touch interactions ($N = n_{Tap} = 13$) while driving with an active ACC ($a_{acc} = 1$) at a speed of 70 km/h. If the model would only access this information, it would predict a TGD of roughly 13 s However, as all interactions were very close to each other ($d_{avg} = 18.224$ px) and were all performed on the homebar ($n_{Homebar} = 13$ without any list or button interaction interfering ($n_{List} = 0$, $n_{Button} = 0$, the final model output is only slightly higher than the average TGD prediction.

These local explanations show that not all features are always relevant. Predictions for secondary task engagements can be driven by only a few dominant features. The presented explanations enable designers and researchers to quickly identify the main forces behind individual predictions. It also allows them to play around with artificial input samples and observe how certain changes in the design of a user flow or the driving situation impact the model's output.

4.3.2. Global explanations

To understand how the features affect the model's output on a global scale, we combine all local explanations of the dataset. Fig. 4 shows the distribution of SHAP values (i.e., the impact of each feature on a specific prediction as seen in 4.3.1) as a set of beeswarm plots. Each dot in a row corresponds to an individual secondary task engagement. The position on the x-axis represents the effect of the respective feature on the model's output. In Fig. 4(a), the SHAP values are in probability space, and in Fig. 4(b) they represent the impact in milliseconds. The color indicates the feature value (red is high, blue is low). The features are sorted by their global importance and only the 19 most important features are displayed individually.

The most important features of the long glance prediction model (Fig. 4(a)), are the number of interactions N, the average distance between the interactions d_{avg}, and the number of tap gestures n_{Tap}. The more touchscreen interactions a driver performs and the larger the distance between them, the higher the output probability that one of the associated glances is longer than 2 s. Fig. 4(a) also reveals that both, the activation of ACC a_{acc} and SA a_{sa}, increase the long glance probability. Whereas the impact of a deactivated assistance system (blue) is small for all samples, the impact varies if the assistance systems are active. The horizontal spread suggests that the impact of assisted driving on visual attention allocation is situation-specific and depends on further factors like the driving situation and interaction patterns. The distribution that describes the impact of the vehicle speed v_{avg} is heavily tailed. For most secondary task engagements at medium speed, the effect is negative but rather small. High speed values reduce the predicted long glance probability and low speed values increase it, respectively. This indicates drivers' self-regulative behavior.

The number of list interactions n_{List} is the most important feature associated with a specific UI element followed by the number of interactions with the homebar $n_{Homebar}$. Through Fig. 4(a), we see that their impact is opposite to each other. Whereas the long glance probability increases with an increasing number of list interactions, it decreases for an increasing number of homebar interactions. This suggests that list interactions tend to be more distracting than interactions on the static homebar. The impact of interactions with Android Auto or Apple Car Play $n_{RemoteUI}$ is similar to the impact of list interactions. In general, we can observe that most of the SHAP value distributions associated with a specific class of UI elements are centered around zero with long tails to one or both sides. This is because most of the elements occur in only a small portion of secondary task engagements. Whereas this leads to a relatively low global importance, these features still have a large impact on specific predictions.

For the TGD prediction model (Fig. 4(b)), N and d_{avg} are also the two most important features. Their distributions also show similarities to the distributions observed in the long glance prediction task. However, the impact of the vehicle speed v_{avg} is inverse compared to the long glance prediction task. High speed values increase the TGD prediction and low values decrease the prediction. Both findings together could be an indication that drivers reduce their single glance duration at higher speeds, which in turn results in longer TGDs because more individual glances are required to complete the same task.

Further, we can see that there are almost no negative contributions associated with UI interaction features. This is due to the fact that the TGD task is cumulative, and every interaction inevitably implies a certain amount of visual attention. However, homebar interactions $n_{Homebar}$, can negatively affect the model output. In line with the observations made for the long glance prediction, list interactions n_{List}, map interactions n_{Map} and interactions with Android Auto and Apple CarPlay $n_{RemoteUI}$ can be associated with an increased visual demand

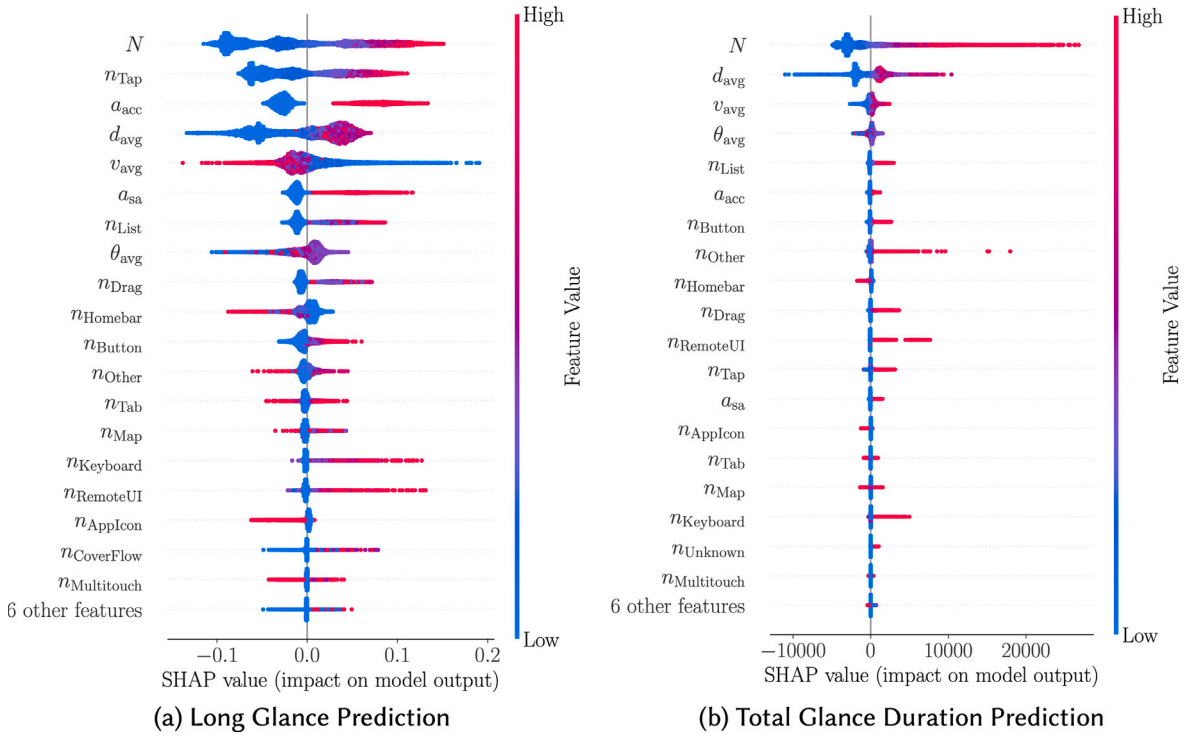


Fig. 4. Explanation summary visualized as a set of beeswarm plots. Each beeswarm plot represents the distribution of SHAP values for one feature.

prediction. A comparison between Figs. 4(a) and 4(b) also reveals that the TGD is not as dependent on the status of the driver assistance systems as the long glance probability.

To understand the effect of a single feature on the model's output in more detail, we plot the SHAP values (y-axis) against the corresponding feature values (x-axis). Every secondary task engagement in our dataset is represented as a dot (see Figs. 5 and 6). Vertical dispersion at a single value on the x-axis shows that there are non-linear dependencies between the displayed feature and other features. To highlight the interaction between features, each dot is colored by the value of the feature that shows the strongest interaction. The histogram at the bottom of the plots shows the distribution of datapoints. Fig. 5(a) suggests that the use of ACC leads to an increased long glance probability prediction. The interaction with the vehicle speed shows that the effect tends to increase with increasing vehicle speed. On the other hand, the data shown in Fig. 5(b) indicates that drivers tend to increase their single glance durations at lower speeds (below 50 km/h) and decrease them at higher speeds (above 125 km/h). However, in between those values, the speed has almost no influence on the model output. This suggests that drivers self-regulate their visual attention allocation based on what they consider an appropriate speed. Additionally, the interaction with the ACC status shows that the impact of the speed on the model output decreases when ACC is active. The interaction effect with a_{acc} partially explains the variance (vertical diversion) in the effect of the vehicle speed. However, various factors like road type or speed limit that may also influence how the vehicle speed affects drivers' visual attention allocation are not considered in the presented models.

Fig. 5(c) indicates that the number of interactions is positively correlated with the drivers' probability to perform a long glance. On the other hand, Fig. 5(d) suggests that as soon as the distance between the touch interactions exceeds a certain threshold (roughly 200 px), the effect on the long glance probability remains constant. Whereas homebar interactions decrease the probability of the model predicting a long glance (Fig. 5(e)), list interactions (Fig. 5(f)) push the model toward predicting a long glance. The interaction effect with the number of interactions additionally indicates that the impact of both elements becomes larger the higher their proportion within a sequence is.

Fig. 6 visualizes how the different features affect the TGD prediction. While the number of interactions N (Fig. 6(c)) is the dominant feature it is also the feature with the highest interaction effect on all other features. Compared to Fig. 5 and in line with the observations we made in Fig. 4, we see that the ACC status a_{acc} (Fig. 6(a)) and the vehicle speed v_{avg} (Fig. 6(b)) do not influence the TGD prediction as much as they influence the long glance prediction. This applies in particular to secondary task engagements with few interactions. The impact of list interactions n_{List} and homebar interactions $n_{Homebar}$ on the TGD, however, is similar to the impact those interactions have on the long glance probability (Figs. 6(e) and 6(f)). This also applies to the influence of the average touch distance d_{avg}. An increase in touch distance leads to an increase in TGD and long glance probability until a certain threshold is reached. However, the interaction effect with N is higher for the TGD prediction model. Another interesting aspect that might need further exploration is the location of the x-intercept. This point describes the touch distance at which the feature's impact turns from decreasing to increasing the visual demand prediction (Fig. 6(d)).

5. Discussion

The presented approach enables users to evaluate the visual demand of early-stage prototypes. In the following, we put our results into perspective and show that the presented approach is more accurate than comparable methods. The predictions and explanations facilitate the generation of fast insights without requiring expensive and long-planned user studies. We illustrate this by assessing three exemplary research objectives covered in the literature. Finally, we address several limitations that apply to our approach.

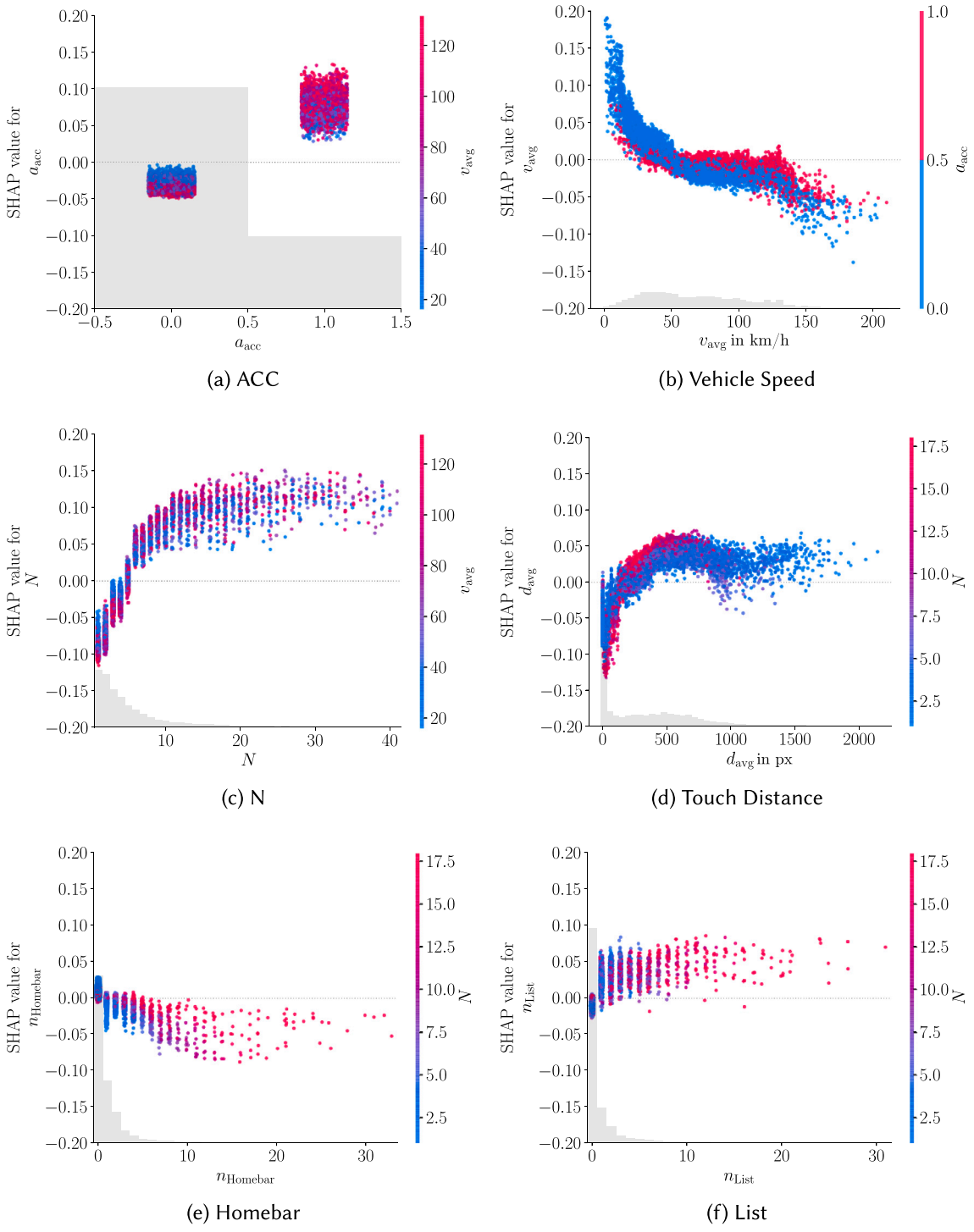


Fig. 5. Feature dependence plots for the long glance classification model.

5.1. *Predicting the visual demand of in-vehicle touchscreen interactions*

Given the complexity of the modeling task, the presented results show how machine learning methods can be used to generate valuable insights into drivers' multitasking behavior by leveraging large naturalistic driving data. Compared to the approach of Kujala and Salvucci (2015), who report critical differences between model predictions and observations, our approach is not only more accurate but also considers a more diverse set of UI elements.

Our approach can predict the TGD with a mean absolute error of roughly 2.4 s over a diverse range of interactions and driving scenarios. In comparison, Purucker et al. (2017) report a mean error of 4 s when

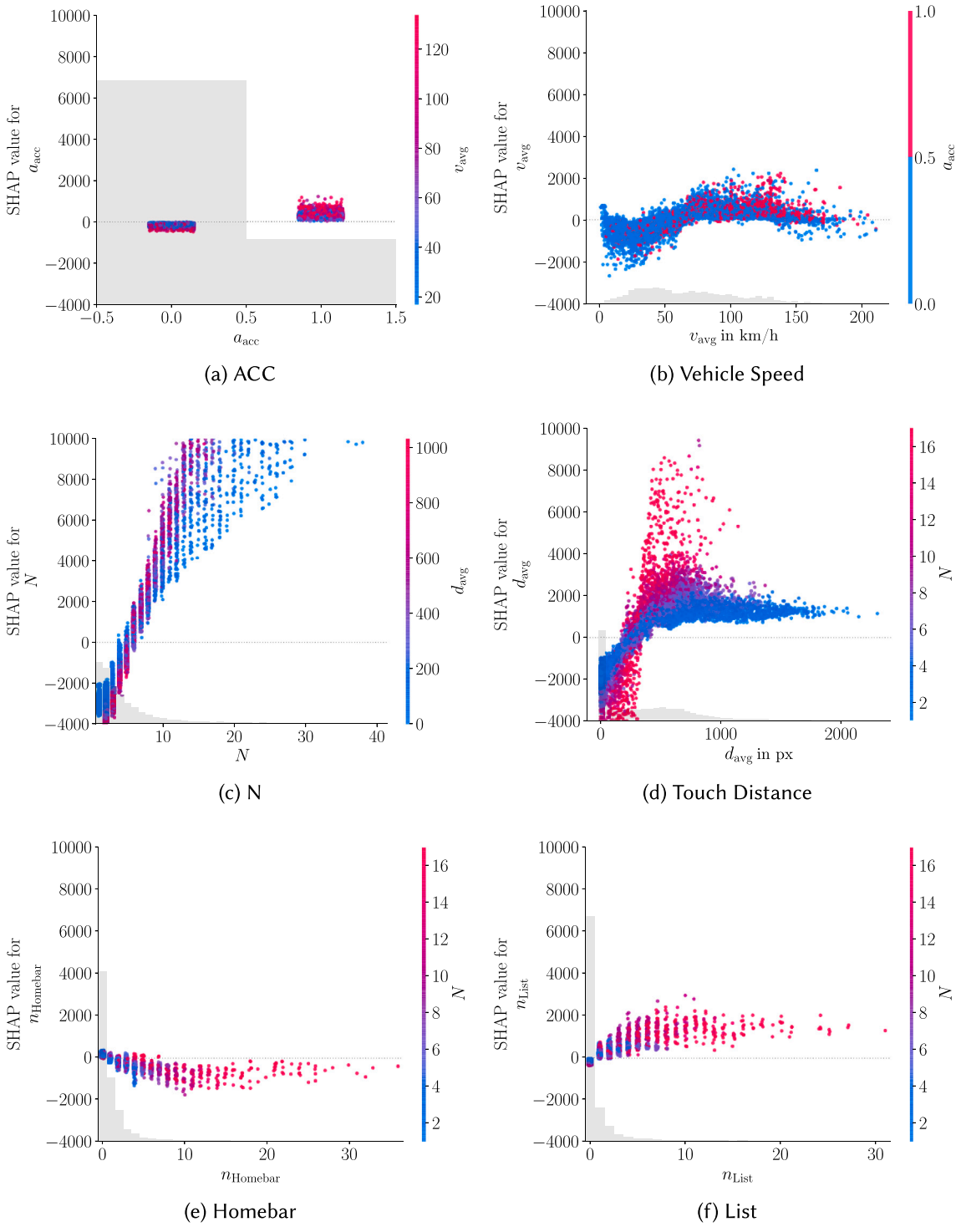


Fig. 6. Feature dependence plots for the TGD model.

averaged over all evaluated tasks. Furthermore, Purucker et al. (2017) use a simple car following task at a constant speed for evaluation. Although these comparisons are useful to put the results into perspective, one needs to consider that the approaches highly differ in their environments and scenarios as described by Janssen et al. (2020).

5.2. *Fast and easily accessible insights based on real-world driving data*

Our approach has two main advantages over conventional user studies. First, the models allow making predictions for yet unseen secondary task engagements. Conventional studies can only be used

to evaluate situations that were explicitly tested. Second, if the user interface undergoes disruptive changes (e.g., a completely new design or concept), the results of a user study are no longer valid and a new study needs to be conducted. Similarly, our computational models may also lose their capability to generalize. However, the advantage of the automated approach for data collection and modeling is that, as soon as a new version is deployed to test vehicles, data is collected and new models based on this new version can be fitted. To demonstrate that our approach is a meaningful extension of traditional user research methods, we compare our results with those from conventional user studies.

The influence of vehicle speed on drivers' visual attention allocation. Risteska et al. (2021) found that an increase in speed reduces drivers' long off-path glances. They argue that drivers modulate their visual attention allocation based on driving demands. A similar finding is presented by Tivesten and Dozza (2014), who found a significant correlation between vehicle speed and off-road glance duration when drivers were engaged in a visual manual phone task. This is consistent with our results shown in Figs. 4(a) and 5(b). Our explanations do not only indicate that the long glance probability decreases with increasing speed but further suggest that this behavior might not be strictly proportional and is also affected by the status of driver assistance systems. Our results further show that the predicted TGD increases with increasing speed (see Fig. 6(b)). The combination of both findings provides a more comprehensive picture suggesting that drivers reduce their single glance duration at higher speeds, forcing them to look to the center stack touchscreen more often. This, in turn, leads to increased TGDs because certain aspects of human glance behavior like the time needed to locate an item are constant for each glance (Large et al., 2017b).

The influence of driving automation on drivers' visual attention allocation. Assisted driving is associated with an increase in the mean and total glance duration during secondary task engagements (Large et al., 2017a; Carsten et al., 2012; Ebel et al., 2022). This is in line with our findings presented in Fig. 4. In a driving simulator study, Carsten et al. (2012) also found that the effect of lateral control (SA) on driver engagement is larger than the effect of longitudinal control (ACC). Based on our data, we cannot confirm this finding. The reasons for this can be manifold but may well be due to the difference between real data and simulation data. Our results further differ from those of Morando et al. (2019), who report no differences in the aggregate off-path glance duration distributions between manual and assisted driving. They only report an effect concerning the on-road glance distribution but state that their eye-tracker did not provide detailed information about the off-path AOIs. Since we can explicitly detect glances toward the center stack touchscreen and can distinguish them from general off-path glances, we argue that our results are superior.

The influence of design characteristics on drivers' visual attention allocation. There are not yet many approaches that have investigated the influence of design characteristics on visual demand in such detail (element type basis) as we show in our approach. Kujala and Salvucci (2015) found that the average distance between two consecutive touch interactions is a critical factor associated with long glances exceeding the limit considered safe. This is in line with our results presented in Figs. 5(d) and 6(d). The explanations that our method provides could additionally serve as a first attempt to quantify the impact spatial separation of interaction elements has on visual demand while driving. Our approach also allows us to make detailed statements about the influence of individual elements. So far, only the task interaction times have been studied in the literature in a roughly similar level of detail (Green et al., 2015; Schneegaß et al., 2011). We found that in particular interactions with maps, lists, and interactions within Apple CarPlay and Android Auto seem to be visually demanding. Interactions on the static homebar, with app icons, and general buttons, on the other hand, are less demanding.

5.3. Benefits for the design process of IVISs and implications on distracted driving prevention

To develop IVISs that are safe to use, driver distraction evaluation needs to be an integral part already in the early design stages. However, driver distraction is a complex construct, and automotive UX experts need data-driven support to evaluate and compare design alternatives concerning their distraction potential (Ebel et al., 2021b). Thus, our approach aims to inform the design process of IVISs from the bottom up to develop solutions that are the least distracting and safe by design. We envision our method to be used to dynamically evaluate early-stage IVIS designs. Users can assess hypothetical IVIS designs concerning their distraction potential in terms of visual demand. They can play around with artificial input samples to learn how changes in the user flow or driving scenario affect drivers' visual attention allocation. Our method then explains how each parameter contributes to the overall prediction. Thus, designers can better understand the effects of various UI elements, driving automation, and vehicle speed on driver distraction. This information can then be used to design IVISs that are less distracting and reduce the risk of accidents. The improved accuracy over comparable approaches and the three application examples show that our approach can make a major contribution to better understanding the complex construct of driver distraction and drivers' visual attention allocation during secondary touchscreen tasks.

5.4. Limitations and future work

As we leverage already commercialized technologies of our research partner, we collected a large amount of behavioral data. We observed drivers' natural interaction behavior without explicitly telling them which touchscreen interactions to perform and therefore eliminate the so-called instruction effect (Carsten et al., 2017). While this approach has many advantages, especially over simulator and test track studies, several limitations apply. These limitations and their potential implications are discussed in the following.

Only company internal cars contributed to the data collection. Whereas they are used for a diverse range of testing procedures, they are also used for transfer and leisure rides of employees, for example over the weekend. We argue that, even if drivers follow a test protocol that aims to evaluate driving-related functions, the incentive to interact with the IVISs does not deviate much from real-world behavior. Furthermore, all drivers in this study need to be considered expert users. However, it is not yet entirely clear to what extent the gaze behavior of experts differs from regular users. Whereas Wikman et al. (1998) report that experienced drivers allocated their visual attention more adequately (Wikman et al., 1998), Naujoks et al. (2016) show that experienced users of Advanced Driver Assistance Systems (ADASs) tend to increase their secondary task engagements compared to novice users. However, a comparison with related approaches (Gaspar and Carney, 2019; Morando et al., 2019; Noble et al., 2021) shows high agreement in total and average glance behavior. Still, the restricted sample of drivers and the fact they were driving alone, need to be considered when interpreting the results.

It is important to consider that the features used in this work do not capture all factors that influence drivers' visual attention allocation. In this study, we only consider the level of driving automation, vehicle speed, and the steering wheel angle to describe the driving situation. These features and their interactions provide valuable information (compare Figs. 5(a), 5(b), 6(b), 6(a), and Fig. A.7 in the Appendix), but they do not allow for a comprehensive description of the driving situation. For example, the effect of vehicle speed may vary not only based on the level of driving automation, but also on the type of road

and traffic situation. Therefore, including additional features may not only improve the description of the driving situation but also make the existing features more meaningful by considering their interaction effects.

Furthermore, it is important to put the results into context and to elaborate on the practical implications this might have. As demonstrated, the approach provides valuable insights into how design artifacts and environmental factors affect drivers' visual attention allocation. The predictions and explanations can guide designers to create interfaces that are less distracting and safer to use. However, even though our approach is superior to related approaches, it is not yet accurate enough to make pixel-precise predictions or to differentiate between minor changes in the driving environment (e.g., driving at 72 km/h vs. 75 km/h). To reliably evaluate the effect of such slight changes or to even act as a basis for driver distraction guidelines, the accuracy needs to be increased. Furthermore, we do not consider environmental factors like lightning conditions or street type (e.g., rural road or highway) or UI artifacts like element color and size that might also influence visual attention allocation. Including such features would provide a more holistic picture and probably more accurate predictions. Moreover, drivers tend to self-regulate their willingness to engage in secondary tasks based on the driving task demands (Ebel et al., 2022; Oviedo-Trespalacios et al., 2018; Hancox et al., 2013). As a result, some interactions occur less frequently in certain driving situations, leading to fewer training data. Therefore, it is likely that prediction accuracy varies across driving situations.

The presented explanations do not imply causality, and therefore do not represent a complete assessment of drivers' visual attention allocation while being engaged in a secondary touchscreen task. However, the explanations help designers to identify the most informative relationships between input features and model outputs, which assist them in understanding the visual demand predicted by the machine learning model.

Having shown that this method delivers promising results, the main goal of future iterations is to improve prediction accuracy. First, a more holistic description of the driving situation by providing additional features like lighting conditions, the proximity of surrounding road users, or map data might lead to significant improvements. Second, considering user demographics like age or driving experience might also lead to better accuracy. Finally, a larger dataset is not only likely to benefit the algorithms presented in this work, but would also enable more sophisticated approaches like recurrent neural networks that can capture sequential information embedded in the interaction sequences.

6. Conclusion

In this paper, we propose a machine learning approach that predicts the visual demand of secondary touchscreen interactions while driving, according to the type of interactions that are performed and the associated driving parameters. Our approach generates local and global explanations providing insights how design artifacts and driving parameters affect drivers' visual attention allocation. We evaluate the approach on a real-world driving dataset consisting of 12,142 secondary task engagements. Our best model identifies secondary task engagements during which drivers perform a long glance with 68% accuracy and predicts the TGD with a mean deviation of 2.4 s. The analysis of the generated explanations reveals clear differences between the visual demand of specific touchscreen interactions and shows that drivers' visual attention allocation depends on the driving situation. In line with related research (Risteska et al., 2021; Tivesten and Dozza, 2014), we show that drivers modulate their visual attention allocation based on the vehicle speed and the level of driving automation.

Our key contributions address many points that previous approaches (Risteska et al., 2021; Large et al., 2017b; Janssen et al., 2015; Victor et al., 2014) have identified as desirable: (1) The approach leverages continuously collected large-scale real-world data providing realistic predictions of drivers' visual attention allocation during secondary task engagements. (2) The approach can easily be adjusted to incorporate additional features and to predict various metrics in addition to TGD and long glance probability (e.g., number of glances, total eyes off-road time, mean glance duration). (3) The local and global explanations provide detailed insight into the impact design artifacts and scenario parameters have on driver distraction prediction. (4) The approach can inform designers about potential implications their design may have and can guide them to design in-vehicle touchscreen interfaces that are safe to use.

CRediT authorship contribution statement

Patrick Ebel: Conceptualization, Methodology, Software, Data curation, Formal analysis, Visualization, Writing – original draft, Writing – review & editing. **Christoph Lingenfelder:** Writing – review & editing, Resources. **Andreas Vogelsang:** Supervision, Writing – review & editing, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Christoph Lingenfelder is an employee of the MBition GmbH which is a subsidiary of Mercedes-Benz. The data used in this work was collected from Mercedes-Benz cars.

Data availability

The authors do not have permission to share data.

Appendix A

A.1. Definitions

Definition 1. An *interaction sequence* $I = (i_n)_{n=1}^{N}$ is a sequence of touchscreen interactions recorded during one trip, where i_n is a single touchscreen interaction performed by a user and N denotes the number of interactions of I. A touchscreen interaction $i = (t, e, p, c)$ is composed of its timestamp t, element type e, gesture type p and coordinate pair $c = (x, y)$. Within I, the duration between two successive interactions $t(i_{n+1}) - t(i_n)$ must be smaller than Δt_{max} such that $t(i_{n+1}) - t(i_n) \leq \Delta t_{max}$.

Definition 2. A *glance sequence* $G = (g_n)_{n=1}^{N}$ is a sequence of non-overlapping intervals of driver glances, where g_n is a single glance performed by a user and N denotes the number of glances of G. Each glance $g_n = (t^s, t^e, r)_n$ is composed of its start time t^s, end time t^e, and AOI r, describing where looked at between t^s and t^e. For all glances of a but the first of a trip, the start time is equal to the end time of the preceding glance $t_s(g_n) = t_s(g_{n-1})$.

Definition 3. A *driving sequence* $D = (d_n)_{n=1}^{N}$ is a sequence of driving data observations, where d_n is a single observation and N denotes the number of observations of D. Each observation is defined as $d_n = (t, v, \theta, a_{ACC}, a_{SA})_n$, where t represents the timestamp, v the vehicle

speed, θ the steering wheel angle, a_{ACC} and a_{SA} the status of the ACC and SA respectively.

Definition 4. A *secondary task engagement* S is defined as an interaction sequence and its corresponding glance sequence and driving sequence $S = (I, G, D)$. We consider all driving observations starting before the first interaction until after the last interaction such that $t(i_1) - t_b < t(d_n) < t(i_N) + t_b$. Where t_b represents a buffer duration. Regarding the glance sequence G, we consider all glances whose start time or end time falls in between the first and last interaction of I such that $t(i_1) < t^s(g_n) < t(i_N) \vee t(i_1) < t^e(g_n) < t(i_N)$.

Problem 1. We define the *long glance prediction task* as the problem of identifying all secondary task engagements S in which a long glance occurs such that for any $g_n \in G$, $\Delta t_g = t^e(g_n) - t^s(g_n) > 2\,\text{s}$ given the according interaction sequence s^i and driving sequence s^d

Problem 2. We define the *total glance duration prediction task* as the problem of predicting the TGD toward the center stack touchscreen during a secondary task engagement S.

A.2. Hyperparameter optimization

In the following we report the results of the hyperparameter optimization for each of the individual models.

A.2.1. Random forest models

The Implementation and the descriptions are based on the scikit-learn python package. For the sets of best performing parameters please refer to Table 3.

n_estimators = [100, 200, 400, 800, 1200, 1600, 2000] — The number of trees in the forest.

max_features = [‘auto’, ‘sqrt’] — Number of features to consider when looking for the best split

max_depth = [10, 20, 30, 40, 60, 80, 100] — Maximum depth of the tree.

min_samples_split = [2, 5, 10] — Minimum number of samples required to split an internal node.

min_samples_leaf = [1, 2, 4] — Minimum number of samples required to be at a leaf node.

bootstrap = [True, False] — Whether bootstrap samples are used when building trees.

Table 3
Sets of best performing parameters for the Random Forest models.

Feature	Long Glance Prediction	TGD Prediction
n_estimators	200	1600
max_features	auto	auto
max_depth	10	60
min_samples_split	5	2
min_samples_leaf	2	4
bootstrap	True	True

A.2.2. XGBoost models

The Implementation and the descriptions are based on the XGBoost python package. For the sets of best performing parameters please refer to Table 4.

n_estimators = [20, 100, 500, 1000, 5000, 10000, 20000] — Number of boosting rounds.

subsample = [0.2, 0.4, 0.6, 0.8, 1] — Subsample ratio of the training instance.

max_depth = [5, 10, 50, 100] — Maximum tree depth for base learners.

learning_rate = [0.0005, 0.001, 0.01, 0.1, 1] — Boosting learning rate (xgb’s “eta”)

colsample_bytree = [0.2, 0.4, 0.6, 0.8, 1] — Subsample ratio of columns when constructing each tree.

colsample_bylevel = [0.2, 0.4, 0.6, 0.8, 1] — Subsample ratio of columns for each level.

Table 4
Sets of best performing parameters for the XGBoost models.

Feature	Long Glance Prediction	TGD Prediction
n_estimators	5000	5000
subsample	0.6	0.8
max_depth	10	10
min_child_weight	4	10
learning_rate	0.01	0.0005
colsample_bytree	0.2	0.6
colsample_bylevel	0.2	1

A.2.3. Feedforward neural networks

The Implementation and the hyperparameter optimization was performed using the Keras API. It needs to be noted that the different hyperparameter combinations did not show large differences in their predictive performance. For the sets of best performing parameters please refer to Table 5.

n_hidden_layers = [1, 2, 3, 4, 5] — Number of layers between the input and output layer of the neural network

n_neurons = [32, 64, 128, 256, 512]) — Number of neurons per layer.

activation = [“relu”, “sigmoid”] — The activation function of the neurons in the respective layer.

drop_out = [0, 0.1, 0.2, 0.3] — The probability at which random units are set to zero during training.

learning_rate = [0.01, 0.001, 0.0001] — Initial learning rate of the ADAM optimizer.

Table 5
Sets of best performing parameters for the FNN models.

Feature	Long Glance Prediction	TGD Prediction
n_hidden_layers	3	1
learning_rate	0.0001	0.001
n_neurons layer 1	512	512
activation layer 1	sigmoid	relu
drop_out layer 1	0.3	0.1
n_neurons layer 2	64	–
activation layer 2	relu	–
drop_out layer 2	0.1	–
n_neurons layer 3	256	–
activation layer 3	sigmoid	–
drop_out layer 3	0.1	–

A.3. Dataset summary statistics

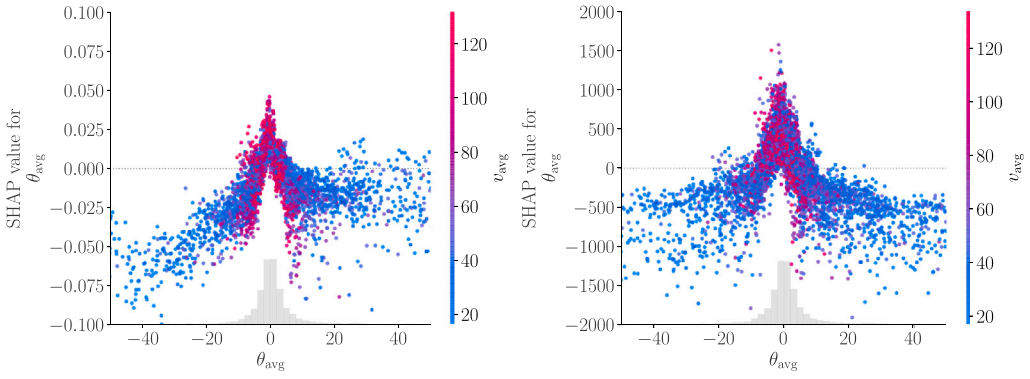
Table 6 provides an overview of all features.

A.4. Steering wheel feature dependence plot

See Fig. A.7.

Table 6
Dataset summary statistics.

Statistic	Mean	St. Dev.	Min	$Q_1(25)$	Median	$Q_3(75)$	Max
Number of interactions	4.431	4.993	1	1	3	5	41
Number of tap gestures	3.814	4.495	0	1	2	5	40
Number of drag gestures	0.363	1.324	0	0	0	0	31
Number of multitouch gestures	0.240	1.108	0	0	0	0	26
Average glance duration in ms	1,441.491	929.736	120.000	960.000	1,241.000	1,659.000	26,801.000
Number of glances	4.367	4.998	1	1	3	6	50
number of long glances	0.569	1.102	0	0	0	1	13
Total glance duration in ms	5,742.751	7,049.487	120.000	1,590.500	3,499.000	7,354.000	262,416.000
average speed in km/h	70.516	36.935	0.633	40.881	66.230	96.567	209.883
ACC active	0.206	0.404	0	0	0	0	1
SA active	0.099	0.299	0	0	0	0	1
AppIcon interactions	0.196	0.586	0	0	0	0	13
CoverFlow interactions	0.038	0.434	0	0	0	0	16
Unknown interactions	0.049	0.450	0	0	0	0	23
Other interactions	0.731	1.568	0	0	0	1	37
List interactions	0.518	1.652	0	0	0	0	31
Tab interactions	0.385	1.335	0	0	0	0	35
ControlBar interactions	0.012	0.135	0	0	0	0	4
Button interactions	0.640	1.515	0	0	0	1	36
Homebar interactions	0.892	2.032	0	0	0	1	36
Slider interactions	0.015	0.228	0	0	0	0	9
ClickGuard interactions	0.058	0.313	0	0	0	0	8
PopUp interactions	0.030	0.232	0	0	0	0	9
Keyboard interactions	0.184	1.435	0	0	0	0	31
Map interactions	0.508	2.181	0	0	0	0	39
RemoteUI interactions	0.173	1.179	0	0	0	0	31
Browser interactions	0.002	0.093	0	0	0	0	8



(a) Long Glance Prediction Sample. (b) Total Glance Duration Prediction Sample.

Fig. A.7. Feature dependence plots of the steering wheel angle and its interaction with the vehicle speed for the long glance classification and TGD model.

References

Anderson, John R., Bothell, Daniel, Byrne, Michael D., Douglass, Scott, Lebiere, Christian, Qin, Yulin, 2004. An integrated theory of the mind. Psychol. Rev. 111 (4), 1036–1060. http://dx.doi.org/10.1037/0033-295x.111.4.1036.

Angell, L.S., Perez, M., Hankey, J., 2008. Driver usage patterns for secondary information systems. In: Invited Paper for the First Human Factors Symposium on Naturalistic Driving Methods & Analyses.

Bärgman, Jonas, Lisovskaja, Vera, Victor, Trent, Flannagan, Carol, Dozza, Marco, 2015. How does glance behavior influence crash and injury risk? A 'what-if' counterfactual simulation using crashes and near-crashes from SHRP2. Transp. Res. F 35, 152–169. http://dx.doi.org/10.1016/j.trf.2015.10.011.

Barredo Arrieta, Alejandro, Díaz-Rodríguez, Natalia, Del Ser, Javier, Bennetot, Adrien, Tabik, Siham, Barbado, Alberto, Garcia, Salvador, Gil-Lopez, Sergio, Molina, Daniel, Benjamins, Richard, Chatila, Raja, Herrera, Francisco, 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion 58, 82–115. http://dx.doi.org/10.1016/j.inffus.2019.12.012.

Burns, Peter, Harbluk, Joanne, Foley, James P., Angell, Linda, 2010. The importance of task duration and related measures in assessing the distraction potential of in-vehicle tasks. In: Proceedings of the 2nd International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI 10. ACM Press, http://dx.doi.org/10.1145/1969773.1969776.

Card, Stuart, 1983. The Psychology of Human-Computer Interaction. L. Erlbaum Associates, Hillsdale, N.J.

Card, Stuart K., Moran, Thomas P., Newell, Allen, 1980. The keystroke-level model for user performance time with interactive systems. Commun. ACM 23 (7), 396–410. http://dx.doi.org/10.1145/358886.358895.

Carsten, O., Hibberd, D., Bärgman, J., Kovaceva, J., Pereira Cocron, M.S., Dotzauer, M., Utesch, F., Zhang, M., Stemmler, E., Guyonvarch, L., Sagberg, F., Forcolin, F., 2017. UDRIVE Deliverable 43.1, Driver Distraction and Inattention, of the EU FP7 Project UDRIVE, first ed. UDRIVE Consortium, BE.

Carsten, Oliver, Lai, Frank C.H., Barnard, Yvonne, Jamson, A. Hamish, Merat, Natasha, 2012. Control task substitution in semiautomated driving. Hum. Factors: J. Hum. Factors Ergonomics Soc. 54 (5), 747–761. http://dx.doi.org/10.1177/0018720812460246.

Custer, Kenny, 2018. 100-Car Data. http://dx.doi.org/10.15787/VTT1/CEU6RB.

de Winter, Joost C.F., Happee, Riender, Martens, Marieke H., Stanton, Neville A., 2014. Effects of adaptive cruise control and highly automated driving on workload and situation awareness: A review of the empirical evidence. Transp. Res. F 27, 196–217. http://dx.doi.org/10.1016/j.trf.2014.06.016.

Dingus, Thomas A., Guo, Feng, Lee, Suzie, Antin, Jonathan F., Perez, Miguel, Buchanan-King, Mindy, Hankey, Jonathan, 2016. Driver crash risk factors and prevalence evaluation using naturalistic driving data. Proc. Natl. Acad. Sci. 113 (10), 2636–2641. http://dx.doi.org/10.1073/pnas.1513271113.

Dingus, T.A., Klauer, S.G., Neale, V.L., Petersen, A., Lee, S.E., Sudweeks, J., Perez, M.A., Hankey, J., Ramsey, D., Gupta, S., Bucher, C., Doerzaph, Z.R., Jermeland, J., Knipling, R.R., 2006. The 100-Car naturalistic driving study: Phase II - results of the 100-Car field experiment. http://dx.doi.org/10.1037/e624282011-001.

Donmez, Birsen, Boyle, Linda Ng, Lee, John D., 2010. Differences in off-road glances: Effects on Young drivers' performance. J. Transp. Eng. 136 (5), 403–409.

Doshi-Velez, Finale, Kim, Been, 2017. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

Ebel, Patrick, Berger, Moritz, Lingenfelder, Christoph, Vogelsang, Andreas, 2022. How do drivers self-regulate their secondary task engagements? the effect of driving automation on touchscreen interactions and glance behavior. In: Proceedings of the 14th International Conference on Automotive User Interfaces and Interactive Vehicular Applications. ACM, Seoul Republic of Korea, pp. 263–273. http://dx.doi.org/10.1145/3543174.3545173.

Ebel, Patrick, Brokhausen, Florian, Vogelsang, Andreas, 2020. The role and potentials of field user interaction data in the automotive UX development lifecycle: an industry perspective. In: 12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications. ACM, Virtual Event DC USA, pp. 141–150. http://dx.doi.org/10.1145/3409120.3410638.

Ebel, Patrick, Lingenfelder, Christoph, Vogelsang, Andreas, 2021a. Visualizing event sequence data for user behavior evaluation of in-vehicle information systems. In: 13th International Conference on Automotive User Interfaces and Interactive Vehicular Applications. In: AutomotiveUI '21, Association for Computing Machinery, New York, NY, USA, pp. 1–10. http://dx.doi.org/10.1145/3409118.3475140.

Ebel, Patrick, Orlovska, Julia, Hünemeyer, Sebastian, Wickman, Casper, Vogelsang, Andreas, Söderberg, Rikard, 2021b. Automotive UX design and data-driven development: Narrowing the gap to support practitioners. Transp. Res. Interdiscip. Perspect. 11, 100455. http://dx.doi.org/10.1016/j.trip.2021.100455.

Engström, Johan, Johansson, Emma, Östlund, Joakim, 2005. Effects of visual and cognitive load in real and simulated motorway driving. Transp. Res. F 8 (2), 97–120. http://dx.doi.org/10.1016/j.trf.2005.04.012.

Fisher, Donald L. (Ed.), 2011. Handbook of Driving Simulation for Engineering, Medicine, and Psychology. CRC Press, Boca Raton.

Fitts, Paul M., 1954. The information capacity of the human motor system in controlling the amplitude of movement. J. Exp. Psychol. 47 (6), 381–391. http://dx.doi.org/10.1037/h0055392.

Gaspar, John, Carney, Cher, 2019. The effect of partial automation on driver attention: A naturalistic driving study. Hum. Factors: J. Hum. Factors Ergonomics Soc. 61 (8), 1261–1276. http://dx.doi.org/10.1177/0018720819836310.

Green, Paul, 1999. Visual and task demands of driver information systems. Technical Report, The University of Michigan Transportation Research Institute (UMTRI).

Green, P., Kang, T., Lin, Brian, 2015. Touch screen task element times for improving SAE recommended practice J2365: First proposal.

Hancox, Graham, Richardson, John, Morris, Andrew, 2013. Drivers' willingness to engage with their mobile phone: The influence of phone function and road demand. IET Intell. Transp. Syst. 7 (2), 215–222. http://dx.doi.org/10.1049/iet-its.2012.0133.

Horrey, William J., Wickens, Christopher D., 2007. In-vehicle glance duration. Transp. Res. Rec. 2018 (1), 22–28. http://dx.doi.org/10.3141/2018-04.

Hutchinson, T.E., White, K.P., Martin, W.N., Reichert, K.C., Frey, L.A., 1989. Human-computer interaction using eye-gaze input. IEEE Trans. Syst. Man Cybern. 19 (6), 1527–1534. http://dx.doi.org/10.1109/21.44068.

ISO15007, 2020. Road vehicles — Measurement and analysis of driver visual behaviour with respect to transport information and control systems. Standard, 2000, International Organization for Standardization, Geneva, CH.

Janssen, Christian P., Boyle, Linda Ng, Ju, Wendy, Riener, Andreas, Alvarez, Ignacio, 2020. Agents, environments, scenarios: A framework for examining models and simulations of human-vehicle interaction. Transp. Res. Interdiscip. Perspect. 8, 100214. http://dx.doi.org/10.1016/j.trip.2020.100214.

Janssen, Christian P., Gould, Sandy J.J., Li, Simon Y.W., Brumby, Duncan P., Cox, Anna L., 2015. Integrating knowledge of multitasking and interruptions across different perspectives and research methods. Int. J. Hum.-Comput. Stud. 79, 1–5. http://dx.doi.org/10.1016/j.ijhcs.2015.03.002.

Kanaan, Dina, Ayas, Suzan, Donmez, Birsen, Risteska, Martina, Chakraborty, Joyita, 2019. Using naturalistic vehicle-based data to predict distraction and environmental demand. Int. J. Mob. Hum. Comput. Interact. 11 (3), 59–70. http://dx.doi.org/10.4018/ijmhci.2019070104.

Kaptein, Nico A., Theeuwes, Jan, van der Horst, Richard, 1996. Driving simulator validity: Some considerations. Transp. Res. Rec. 1550 (1), 30–36. http://dx.doi.org/10.1177/0361198196155000105.

Klauer, Sheila, Dingus, Thomas, Neale, T, Sudweeks, J., Ramsey, D, 2006. The impact of driver inattention on near-crash/crash risk: An analysis using the 100-Car naturalistic driving study data. Technical Report, 594, U.S. Department of Transportation, National Highway Traffic Safety Administration / Virginia Tech Transportation Institute, 3500 Transportation Research Plaza (0536) Blacksburg, Virginia 24061.

Kohavi, Ron, 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2. In: IJCAI'95, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 1137–1143.

Kujala, Tuomo, Salvucci, Dario D., 2015. Modeling visual sampling on in-car displays: The challenge of predicting safety-critical lapses of control. Int. J. Hum.-Comput. Stud. 79, 66–78. http://dx.doi.org/10.1016/j.ijhcs.2015.02.009.

Kutila, Matti, Jokela, Maria, Markkula, Gustav, Rue, Maria Romera, 2007. Driver distraction detection with a camera vision system. In: 2007 IEEE International Conference on Image Processing. IEEE, http://dx.doi.org/10.1109/icip.2007.4379556.

Large, David, Banks, Victoria, Burnett, Gary, Baverstock, Sarah, Skrypchuk, Lee, 2017a. Exploring the behaviour of distracted drivers during different levels of automation in driving.

Large, David R., Burnett, Gary, Crundall, Elizabeth, van Loon, Editha, Eren, Ayse L., Skrypchuk, Lee, 2017b. Developing predictive equations to model the visual demand of in-vehicle touchscreen HMIs. Int. J. Hum.-Comput. Interact. 34 (1), 1–14. http://dx.doi.org/10.1080/10447318.2017.1306940.

Lee, Seul Chan, Yoon, Sol Hee, Ji, Yong Gu, 2019. Modeling task completion time of in-vehicle information systems while driving with keystroke level modeling. Int. J. Ind. Ergon. 72, 252–260. http://dx.doi.org/10.1016/j.ergon.2019.06.001.

Li, Zhaojian, Bao, Shan, Kolmanovsky, Ilya V., Yin, Xiang, 2018. Visual-manual distraction detection using driving performance indicators with naturalistic driving data. IEEE Trans. Intell. Transp. Syst. 19 (8), 2528–2535. http://dx.doi.org/10.1109/tits.2017.2754467.

Li, Li, Zhong, Boxuan, Hutmacher, Clayton, Liang, Yulan, Horrey, William J., Xu, Xu, 2020. Detection of driver manual distraction via image-based hand and ear recognition. Accid. Anal. Prev. 137, 105432. http://dx.doi.org/10.1016/j.aap.2020.105432.

Liang, Yulan, Lee, John D., 2010. Combining cognitive and visual distraction: Less than the sum of its parts. Accid. Anal. Prev. 42 (3), 881–890. http://dx.doi.org/10.1016/j.aap.2009.05.001.

Liao, Q. Vera, Gruen, Daniel, Miller, Sarah, 2020. Questioning the AI: informing design practices for explainable AI user experiences. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. ACM, Honolulu HI USA, pp. 1–15. http://dx.doi.org/10.1145/3313831.3376590.

Lipton, Zachary C., 2018. The mythos of model interpretability. Queue 16 (3), 31–57. http://dx.doi.org/10.1145/3236386.3241340.

Lundberg, Scott M., Erion, Gabriel, Chen, Hugh, DeGrave, Alex, Prutkin, Jordan M., Nair, Bala, Katz, Ronit, Himmelfarb, Jonathan, Bansal, Nisha, Lee, Su-In, 2020. From local explanations to global understanding with explainable AI for trees. Nat. Mach. Intell. 2 (1), 56–67. http://dx.doi.org/10.1038/s42256-019-0138-9.

Lundberg, Scott M., Erion, Gabriel G., Lee, Su-In, 2018. Consistent individualized feature attribution for tree ensembles. arXiv preprint arXiv:1802.03888.

Lundberg, Scott M., Lee, Su-In, 2017. A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. In: NIPS'17, Curran Associates Inc., Red Hook, NY, USA, pp. 4768–4777.

Manes, Daniel I., 1997. Prediction of destination entry and retrieval times using keystroke-level models.

Merchant, John, 1967. The oculometer. Technical Report NASA-CR-805, National Aeronautics and Space Administration.

Merlhiot, Gaëtan, Bueno, Mercedes, 2021. How drowsiness and distraction can interfere with take-over performance: A systematic and meta-analysis review. Accid. Anal. Prev. 106536. http://dx.doi.org/10.1016/j.aap.2021.106536.

Molnar, Christoph, 2020. Interpretable Machine Learning. Lulu.com.

Morando, Alberto, Gershon, Pnina, Mehler, Bruce, Reimer, Bryan, 2021. Visual attention and steering wheel control: From engagement to disengagement of Tesla Autopilot. Proc. Hum. Factors Ergonomics Soc. Ann. Meet. 65 (1), 1390–1394. http://dx.doi.org/10.1177/1071181321651118.

Morando, Alberto, Victor, Trent, Dozza, Marco, 2019. A reference model for driver attention in automation: Glance behavior changes during lateral and longitudinal assistance. IEEE Trans. Intell. Transp. Syst. 20 (8), 2999–3009. http://dx.doi.org/10.1109/tits.2018.2870909.

Naujoks, Frederik, Purucker, Christian, Neukum, Alexandra, 2016. Secondary task engagement and vehicle automation – Comparing the effects of different automation levels in an on-road experiment. Transp. Res. F 38, 67–82. http://dx.doi.org/10.1016/j.trf.2016.01.011.

NHTSA, 2012. Visual-manual NHTSA driver distraction guidelines for in-vehicle electronic devices. Standard, National Highway Traffic Safety Administration (NHTSA).

Noble, Alexandria M., Miles, Melissa, Perez, Miguel A., Guo, Feng, Klauer, Sheila G., 2021. Evaluating driver eye glance behavior and secondary task engagement while using driving automation systems. Accid. Anal. Prev. 151, 105959. http://dx.doi.org/10.1016/j.aap.2020.105959.

Oviedo-Trespalacios, Oscar, Haque, Md. Mazharul, King, Mark, Washington, Simon, 2018. Should I text or call here? a situation-based analysis of drivers' perceived likelihood of engaging in mobile phone multitasking: mobile phone multitasking engagement. Risk Anal. 38 (10), 2144–2160. http://dx.doi.org/10.1111/risa.13119.

Pampel, Sanna M., Burnett, Gary, Hare, Chrisminder, Singh, Harpreet, Shabani, Arber, Skrypchuk, Lee, Mouzakitis, Alex, 2019. Fitts goes autobahn. In: Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications. ACM, http://dx.doi.org/10.1145/3342197.3344538.

Pettitt, Michael, Burnett, Gary, Stevens, Alan, 2007. An extended keystroke level model (KLM) for predicting the visual demand of in-vehicle information systems. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, http://dx.doi.org/10.1145/1240624.1240852.

Purucker, Christian, Naujoks, Frederik, Prill, Andy, Neukum, Alexandra, 2017. Evaluating distraction of in-vehicle information systems while driving by predicting total eyes-off-road times with keystroke level modeling. Applied Ergon. 58, 543–554. http://dx.doi.org/10.1016/j.apergo.2016.04.012.

Ribeiro, Marco Tulio, Singh, Sameer, Guestrin, Carlos, 2016. Why should I trust you?. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, http://dx.doi.org/10.1145/2939672.2939778.

Riener, Andreas, 2011. Assessment of simulator fidelity and validity in simulator and on-the-road studies. Int. J. Adv. Syst. Meas. 3, 110–124 (15).

Risteska, Martina, Kanaan, Dina, Donmez, Birsen, Chen, Huei-Yen Winnie, 2021. The effect of driving demands on distraction engagement and glance behaviors: Results from naturalistic data. Saf. Sci. 136, 105123. http://dx.doi.org/10.1016/j.ssci.2020.105123.

SAEJ3016, 2021. SAEJ3016: Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. Standard, Society of Automotive Engineers (SAE), Warrendale.

Schneegaß, Stefan, Pfleging, Bastian, Kern, Dagmar, Schmidt, Albrecht, 2011. Support for modeling interaction with automotive user interfaces. In: Proceedings of the 3rd International Conference on Automotive User Interfaces and Interactive Vehicular Applications. In: AutomotiveUI '11, Association for Computing Machinery, New York, NY, USA, pp. 71–78. http://dx.doi.org/10.1145/2381416.2381428.

Shapley, L.S., 1953. A value for N-Person games. In: Contributions to the Theory of Games (AM-28), Volume II. Princeton University Press, pp. 307–318. http://dx.doi.org/10.1515/9781400881970-018.

Shin, Donghee, 2021. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. Int. J. Hum.-Comput. Stud. 146, 102551. http://dx.doi.org/10.1016/j.ijhcs.2020.102551.

Tivesten, Emma, Dozza, Marco, 2014. Driving context and visual-manual phone tasks influence glance behavior in naturalistic driving. Transp. Res. F 26, 258–272. http://dx.doi.org/10.1016/j.trf.2014.08.004.

Tsimhoni, Omer, Green, Paul, 2001. Visual demand of driving and the execution of display-intensive in-Vehicle tasks. Proc. Hum. Factors Ergonomics Soc. Ann. Meeting 45 (23), 1586–1590. http://dx.doi.org/10.1177/154193120104502305.

Verma, Tejaswani, Lingenfelder, Christoph, Klakow, Dietrich, 2020. Defining explanation in an AI context. In: Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP. Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.blackboxnlp-1.29.

Victor, Trent, Dozza, Marco, Bärgman, Jonas, Boda, Christian-Nils, Engström, Johan, Markkula, Gustav, 2014. Analysis of naturalistic driving study data: Safer glances, driver inattention, and crash risk.

Wang, Chao, An, Pengcheng, 2021. A mobile tool that helps nonexperts make sense of pretrained CNN by interacting with their daily surroundings. In: Adjunct Publication of the 23rd International Conference on Mobile Human-Computer Interaction. ACM, Toulouse & Virtual France, pp. 1–5. http://dx.doi.org/10.1145/3447527.3474873.

Wiegand, Gesa, Eiband, Malin, Haubelt, Maximilian, Hussmann, Heinrich, 2020. "I'd like an explanation for that!"exploring reactions to unexpected autonomous driving. In: 22nd International Conference on Human-Computer Interaction with Mobile Devices and Services. ACM, Oldenburg Germany, pp. 1–11. http://dx.doi.org/10.1145/3379503.3403554.

Wikman, Anna-Stina, Nieminen, Tapio, Summala, Heikki, 1998. Driving experience and time-sharing during in-car tasks on roads of different width. Ergonomics 41 (3), 358–372. http://dx.doi.org/10.1080/001401398187080.

Wintersberger, Philipp, Schartmüller, Clemens, Shadeghian-Borojeni, Shadan, Frison, Anna-Katharina, Riener, Andreas, 2021. Evaluation of imminent take-over requests with real automation on a test track. Hum. Factors: J. Hum. Factors Ergonomics Soc. 001872082110514. http://dx.doi.org/10.1177/00187208211051435.

Wollmer, M, Blaschke, C, Schindl, T, Schuller, B, Farber, B, Mayer, S, Trefflich, B, 2011. Online driver distraction detection using long short-term memory. IEEE Trans. Intell. Transp. Syst. 12 (2), 574–582. http://dx.doi.org/10.1109/tits.2011.2119483.

Zhang, Yunfeng, Liao, Q. Vera, Bellamy, Rachel K.E., 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. ACM, Barcelona Spain, pp. 295–305. http://dx.doi.org/10.1145/3351095.3372852.