Towards Mental Imagery-Aware Systems using Physiological Sensors and Machine Learning

Dissertation

Thesis approved by
the Department of Computer Science
University of Kaiserslautern-Landau
for the award of the Doctoral Degree
Doctor of Engineering (Dr.-Ing.)

to

Brishtel Iuliia

Date of Defence:	May 3rd 2024
Dean:	Prof. Dr. Christoph Garth
Supervisor:	Prof. Dr. Didier Stricker
Supervisor:	Prof. Dr. Prof. h.c. Andreas Dengel

DE-386

Abstract

Context-aware systems use their functional environment to provide relevant information or functions to their users. Currently available context-aware systems, for example, advanced driving assistant systems or mobile learning applications, primarily rely on physical activity or direct user input. This thesis makes the next step and introduces the concept of mental-imagery-aware systems that could enable a more sophisticated perception of users. Mental imagery includes multiple dimensions, whereas this thesis focuses on the two most common forms occurring in daily life: mind-wandering and spatial imagery.

Mind-wandering-aware systems are especially relevant in learning settings, where mind-wandering itself is primarily associated with low learning performance. This work proposes a novel approach for detecting episodes of mind wandering using physiological sensors and machine learning methods. For the first time, it is demonstrated that the electrodermal activity is sufficient for classifying the episodes of mind wandering with outstanding classification accuracy. The fusion of eye-tracker and electrodermal activity data additionally improves the classification performance of the machine learning algorithms.

Next, this thesis introduces a prospect towards spatial imagery and engagement-aware systems. With the rapid increase of automation levels in serial vehicles, there is a need for a better understanding of its impact on spatial imagery required for successful navigation. For this purpose, a highly immersive driving simulation system with an integrated eye-tracking system is deployed. With a real-time application in mind, the proxy of spatial imagery "engagement with a driving task" is used to infer the driver's presence in the driving loop. The work demonstrates the feasibility of the eye-tracking features combined with the Gradient boosting algorithm to recognize the disengagement of the driver from the driving loop outperforming the state-of-the-art. Finally, this work pave the wave for the driver's engagement recognition using a UWB-radar and deep learning algorithms. Six driving activities are recorded with a UWB radar and classified using state-of-the-art deep learning models showing promising outcomes.

Acknowledgment

I would like to express my deep gratitude to my dissertation advisor, Prof. Didier Stricker, for providing me with a scientific environment to complete my PhD at the Department of Augmented Vision, DFKI. He encouraged me to push boundaries and step beyond my own shadow. I would like to give special thanks to Prof. Andreas Dengel, who supported my scientific career in Computer Sciences. During my Master's, he recognized my passion for Artificial Intelligence and Japanese culture and gave me the opportunity to work at his lab, "Immersive Quantified Learning" (iQL) at DFKI, Kaiserslautern. I would like to thank Prof. Stefan Deßloch for chairing my viva and Dr. Bernd Schürmann for his support with PhD-related matters.

I also acknowledge Prof. Thomas Schmidt and Prof. Thomas Lachmann. Prof. Schmidt invited me to join his experimental psychology lab during my Bachelor's, laying the first cornerstone of my scientific career. He taught me the best practices of scientific research, and his insightful advice supported me throughout my PhD. Prof. Lachmann's support through a PhD scholarship was invaluable. My special appreciation goes to Dr. Rekrut Maurice, Dr. Tilman Dingler, and Dr. Markus Weber for the research collaborations. I am deeply thankful to all participants who invested their time in taking part in my studies.

I would like to thank Dr. Jason Rambach for including me in his research group and for sharing the necessary resources. Special thanks go to Dr. Nicolas Großmann for providing me with opportunities for research stays at the University of Melbourne and for making our time at iQL so enjoyable. I am very grateful to Christian Schulze for his support in utilizing clusters effectively. My deepest thanks to my former PhD colleagues Stephan, Mahdi, Shoya, and Stephan, as well as my student assistants Emil and Robert, for their incredible technical support and fruitful discussions. I would also like to extend my deepest appreciation to Benjamin for his very special and strong support throughout the long period of my PhD.

I am inexpressibly thankful to my dearest Igor for helping to put into practice many of my scientific ideas, motivating me at every step, and providing unconditional support during challenging times. Lastly, my deepest gratitude to my beloved parents, Oksana and Viktor, and my brother, Daniil, for their patience, belief in me, continuous support, and the opportunity to follow my dreams. This thesis is submitted with the warmest memories of my grandmother Maria, whose kindness, wisdom, and superb work ethic continue to inspire me.

”I have never tried that before, so I think I should definitely be able to do that.”

Pippi Longstocking

Abbreviations

ANOVA	Analysis of Variance
AOI	Area of Interest
CNN	Convolutional Neural Network
CV	Cross Validation
EDA	Electrodermal Activity
HAR	Human Activity Recognition
IQR	Interquartile Range
KNN	K-Nearest Neighbors Algorithm
LOO	Leave-One-Out
LSTM	Long Short-Term Memory
PVT	Pyramid Vision Transformer
RNN	Recurrent Neural Network
SCL	Skin Conductance Level
SCR	Skin Conductance Response
SHAP	Shapley Additive Explanations
SMOTE	Synthetic Minority Over-Sampling Technique
SVM	Support Vector Machine
TRT	Task Related Thoughts
TUT	Task Unrelated Thoughts
UWB	Ultrawide Band

List of Figures

List of Tables

Contents

1 Introduction

The increasing amount of available data and the number of sensors embedded in electronic devices of daily use enhance a comprehensive perception of users, their physical and mental state, intention, and goals. As a result, daily live devices supplied with context-aware applications have grown enormously and spread from mobile devices up to driving assistant systems. Thus, Apple Watches Series 4 can automatically detect a hard fall of its owner and make an emergency call (see Figure 1.1). Advanced driving assistant systems can detect distracted drivers' behavior, trigger a warning signal, or adapt the vehicle's behavior accordingly. Learning assistant systems can provide with more tailored environment enhancing personal feed and elaborating more effective learning strategies. In particular, eye-tracker-equipped learning systems can recognize possible difficulties in topic understanding based on users' gaze and thus adopt the information representation accordingly [119] (see Figure 1.1).

The interest in context-aware systems has been gradually increasing, as the tendency of the last five years shows [64]. This development primarily emerges as a logical step to maximize the use of constantly growing available contextual and sensory data. The goal is to enhance the user experience by tailoring the systems or services to user's needs at the given moment. A system is defined to be context-aware if: *it uses context to provide relevant information and/or services to the user, where relevancy depends on the user's task* [106, 6]. The term *context* itself can be defined as *any information that can be used to characterize the situation of an entity, where the entity is a person, place, or object that is considered relevant to the interaction between a user and its application, including the user and the application themselves* [106].

In the midst of increasing automation and adaptation level of assistant systems, more complex information about users might be required to guarantee the proper system response and an appropriate safety level. Thus, recognising physical activities like the presence of the drivers' hands on the steering wheel does not provide sufficient

Figure 1.1: Example of context-aware systems. **Top left**: Apple Smartwatches Series 4 with a fall detection recognition [3]. **Top right**: Bosch interior monitoring system [28]. **Bottom**: Gaze-aware learning assistant system Hypermind [109].

inference whether they are fully aware of the ongoing road situation [118]. Neither the position of the gaze on a particular part of the text is always revelatory to understanding whether the user perceives the information. Considering these limitations, this thesis focuses on the concept of *mental imagery* that, by now, has not received much attention in context-aware systems.

1.1 Motivation

Mental imagery is the ability of the brain to generate, inspect, and manipulate the internal mental representation of objects, events, and environments that are physically not presented [176]. It is assumed to be a quasi-perceptual experience resembling a perception but without any direct sensory input [208]. The forms and vividness of mental imagery vary from individual to individual, making it a strictly subjective

Figure 1.2: Common forms of mental imagery in daily life. **Left**: Spatial imagery. **Right**: Mind wandering. Credits: Adobe Stock.

experience. Both extremes on the mental imagery scale can indicate mental illnesses and clinical disorders. Thus, the lack of ability to voluntarily experience mental images is called *aphantasia* [162, 79], which can occur both in healthy individuals [121] and those with a brain injury. Aphantasia's counterpart is an involuntary experience of vivid, intensive mental images decoupled from the direct sensory input [162], typical for patients with schizophrenia, depression, and anxiety Parkinson's disease. Apart from both extremes, there is strong evidence that mental imagery is a regular function of the healthy brain. It allows individuals to simulate the future, analyze past events, engage in self-related thoughts, and moderate goal-oriented and planning behavior [136].

Mental imagery is a general term aggregating several cognitive functions, including *mind wandering* and *navigation* or *spatial imagery* [162]. Mind wandering can be defined as attentional decoupling from physical activity to internal thoughts [195]. It was shown to force creativity and problem-solving ability [216, 195], as well as to moderate reading comprehension [202]. The navigation is heavily based on *spatial imagery* - the cognitive ability to inspect and evaluate spatial features (e.g. distance, relative position) of mentally generated images [176]. The planning ability of spatial navigation has a crucial role in everyday life, especially in the context of driving, since a good navigation ability is a crucial component for on-road safety, accident prevention, and pollution and energy consumption [110]. Successful navigation is only possible when individuals are fully engaged in the navigation process: they perceive the

environment, know their position, and can identify the destination point from the current position. Therefore, engagement with driving is further considered as an essential component moderating the strengths of spatial imagery. Considering the prevalent role of mind wandering and spatial imagery in such daily activities as learning and driving, the development of mental imagery-aware systems can significantly enhance the learning and driving experience in the respective domains.

1.2 Existing Challenges

Although the research on mind wandering in the context of learning assistant systems has been experiencing a rapid increase, mental imagery itself is not a primary focus of context-aware systems. Several factors could be accountable for it: (1) the hidden nature of mental imagery, and therefore (2) the absence of stable sampling methods, (3) high subjectiveness, (4) context dependency, and finally, (5) the antagonistic nature of mental imagery to be both detrimental and beneficial. Because of the covert and, in some cases, uncontrolled nature of mental imagery [165], the dominant research tool for mental imagery has been remaining self-reported questionnaires and experience sampling methods. However, both these techniques lack sufficient subjectivity [163]: a questionnaire provides only a proxy representation of experienced imagery and heavily depends on participants' awareness of their thoughts [182]. In addition, a continuous request to report the presence of mental imagery might disturb the communication flow between the user and the system. In contrast, neuro-imagery and brain stimulation techniques can reveal keen insights about the brain activity associated with mental imagery. Nevertheless, their deployment is restricted to clinical settings due to complexity and a limited degree of freedom for users, making them impractical for context-aware applications. Finally, the observation and registration of the mental imagery are typically limited to inference statistics and do not translate to context-aware applications.

1.3 Problem Formulation and Approach

Addressing the aforementioned limitations from Section 1.2, this work aims to investigate the feasibility of low-cost, scalable physiological sensors combined with machine

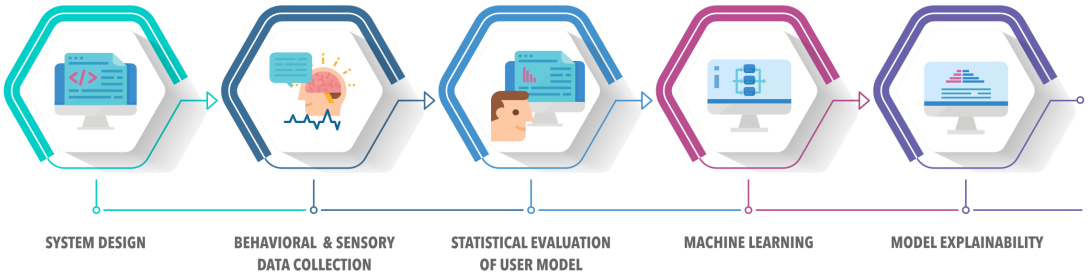


Figure 1.3: Pipeline for mental imagery-aware system development: Five successive steps for a robust system framework.

learning algorithms to provide prospectives towards mental imagery-aware systems. Considering the subjective and hidden manner of mental imagery, this thesis sets a goal to design systems with a strong generalization ability for new, previously unseen users in mental imagery classification tasks.

In this thesis, a mental-imagery aware system is designed as *a system that uses users' data to infer the presence of mental imagery in a particular context*. Considering the need for the precise definition of the system context and multidimensionality of mental imagery itself, the thesis investigates two application scenarios for mental imagery-aware systems, namely *learning* and *driving*, as these two activities are an integral part of daily life. Importantly, as mentioned in Section 1.1, engagement is a key component of spatial image. Therefore, it can provide a proxy of an ongoing spatial imagery process in cases the latter one can not be directly captured by the sensors [176].

The system prototyping for both application scenarios follows the schema introduced in Figure 1.3. The proposed pipeline includes five successive steps: (1) *System Design*, (2) *Behavioral and Sensory Data Collection*, (3) *Statistical Analysis*, (4) *Machine Learning*, and (5) *Model Explainability*. *System Design* introduces a particular architecture and sensors used in the system. In the second step, data containing user responses are collected. These data are further used as ground truth for sensory data. In the third step, *statistical analysis* is performed to test pre-defined hypotheses regarding user behavior and designed context. Next, machine learning models are built to enable the system to correctly classify the presence of the mental imagery by the user. Finally, *model explainability* techniques are used to reveal the feature's impact on the classification per-

formance of the built models. These five steps should ensure a robust and substantiated framework for developing mental imagery-aware systems.

The physiological sensors, namely an eye tracker, an electrodermal activity sensor and a radar used in this thesis for mental imagery-aware systems, were selected according to their scalability, cost, privacy, interchangeability among application scenarios, and performance reliability under learning and driving conditions. Taking into account that failures in visual attention are responsible for a significant number of traffic accidents [118], there is no wonder that eye tracking is widely employed in autonomous driving research and partly integrated into serial vehicles [149, 128, 77]. For the driver monitoring systems, the radar sensor is additionally deployed, as it overcomes the limitations of the camera-based systems, such as sensitivity to the lighting conditions, as well as privacy issues.

Thus, the main research question of this thesis is:

Can machine learning combined with physiological sensors enhance the quantification of mental imagery in context-aware systems?

First, finding sensors that can depict the episodes of mental imagery without intruding into users' privacy and autonomy [106] is of high importance. Second, deploying machine learning techniques for acquired sensory data should enable the system to automatically classify the presence of mental imagery in the defined context. Taking into account the system dependency on a precise context definition, the main research question of this thesis is divided into three subquestions:

1. **Question**: How can the features extracted from the electrodermal activity and eye movement, combined with machine learning, contribute to detecting episodes of mind wandering in a learning context?

2. **Question**: How can machine learning contribute to detecting spatial imagery and engagement with a driving task under various driving conditions?

3. **Question**: How can the radar system contribute to driver-independent, privacy-driven monitoring solutions in the context of engagement-aware systems?

1.4 Contributions

In the following, the key contributions of this thesis are summarized. All contributions have been published in papers or are in preparation, as listed below.

1. This thesis introduces a novel method for user-independent classification of episodes of mind wandering using electrodermal activity, an eye-tracker, and machine learning methods. It demonstrates for the first time that using features extracted from electrodermal activity alone is sufficient for outstanding classification performance. The eye-tracking feature-based model outperformed the state-of-the-art mind wandering classification task. Finally, it provides a pipeline for developing mind wandering-aware assistant systems. This work has been published in:

   > **Brishtel**, I.; Khan, A.A.; Schmidt, T.; Dingler, T.; Ishimaru, S.; Dengel, A. Mind Wandering in a Multimodal Reading Setting: Behavior Analysis & Automatic Detection Using Eye-Tracking and an EDA Sensor. Sensors 2020, 20, 2546, doi.org/10.3390/s20092546.

2. This thesis fills the gap in understanding how autonomous and navigated driving impacts spatial imagery and cognitive maps, informing future advancements in driver-assistance technologies. Thus, the impact of printed maps, navigation systems, and autopilot in a virtual environment was compared to examine the development of spatial knowledge and cognitive demands under various driving conditions. Learning a new route with printed maps was associated with a higher cognitive demand than the navigation system and autopilot. Conversely, driving a route by memory resulted in an increased cognitive workload if the route had been previously learned with the navigation system or autopilot. Way-finding performance was found to be less prone to errors when learning a route from a printed map. This work has been published in:

   > **Brishtel**, I.; Schmidt, T.; Vozniak, I.; Rambach, J.R.; Mirbach, B.; Stricker, D. To Drive or to Be Driven? The Impact of Autopilot, Navigation System, and Printed Maps on Driver's Cognitive Workload and Spatial Knowledge. ISPRS Int. J. Geo-Inf., 2021, doi.org/10.3390/ijgi10100668.

3. This thesis introduces a novel, user-independent approach to classify the presence of the driver in the driving loop based on the gaze data. Built upon the data collected from the study on spatial imagery and driving conditions, this work demonstrates the feasibility of the eye-tracking features combined with the Gradient Boosting algorithm to recognize the disengagement of the driver from the driving loop, even when the eyes are directed on the road. The proposed method of classification performance of the driving mode outperforms the state-of-the-art. This work has been published in:

   > **Brishtel**, I.; Krauss S.; Schmidt, T.; Rambach, J.R.; Vozniak, I.; Stricker, D. Classification of Manual Versus Autonomous Driving based on Machine Learning of Eye Movement Patterns. 2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Prague, Czech Republic, 2022, pp. 700-705, doi.org/10.1109/SMC53654.2022.9945234

4. This thesis shows that drivers' engagement and activity classification can be performed using ultra-wideband radar technology, providing ultimate advantages in terms of privacy and robustness to environmental conditions. It is demonstrated for the first time how a convolutional neural network, a long short-term memory network, and a visual Transformer can facilitate driver activity and engagement recognition using Doppler data.

5. Finally, a novel, previously not existing dataset *RaDa* containing driving activities is introduced and made publicly available for the research community. A total of 10.406 frames are included in the dataset, each with Doppler range information. This work has been published in:

   > **Brishtel**, I.; Krauss, S.; Chamseddine, M.; Rambach, J.R.; Stricker, D. Driving Activity Recognition Using UWB Radar and Deep Neural Networks. Sensors 2023, 23, 818, doi.org/10.3390/s23020818.

1.5 Organization of the Thesis

This thesis is organized as follows: Chapter 2 introduces notations and provides relevant information for understanding the thesis. It includes the description of physiological

sensors used in this work, their working principles and provides background about the physiological mechanisms measured by these sensors. Finally, it discusses the basic idea and mathematical background behind the in this thesis deployed machine learning methods. Chapter 3 provides information about the case study on mind-wandering recognition in a learning setting. It discusses the preprocessing steps, feature engineering, and feature importance for automatic, user-independent mind-wandering classification tasks. Chapter 4 provides a prospective towards developing the engagement-aware system in a driving context. For this purpose, it investigates the user-model and spatial imagery of drivers under various navigation conditions to fill the existing gap in the state-of-the-art. Based on the observed results it offers a prospective to classify spatial imagery through the observed engagement level of the driver while navigating in a virtual city using gaze data and machine learning algorithms. Chapter 5 provides a study addressing the engagement recognition of the driver with the driving task using a radar-based system. Moreover, it introduces a novel dataset, the first of its kind. The radar preprocessing steps are addressed in the chapter, and the classical machine learning algorithm is compared to deep learning methods. Finally, Section 6 concludes the thesis and provides suggestions and outlooks for future work.

1.6 Publications

Most of the work presented in this thesis has been accepted and presented at peer-reviewed conferences. In the following, a list of the papers published during the time of the PhD is provided:

Journals

1. **Brishtel**, I.; Krauss, S.; Chamseddine, M.; Rambach, J.R.; Stricker, D. Driving Activity Recognition Using UWB Radar and Deep Neural Networks. Sensors 2023, 23, 818, doi.org/10.3390/s23020818.

2. **Brishtel**, I.; Schmidt, T.; Vozniak, I.; Rambach, J.R.; Mirbach, B.; Stricker, D. To Drive or to Be Driven? The Impact of Autopilot, Navigation System, and Printed Maps on Driver's Cognitive Workload and Spatial Knowledge. ISPRS Int. J. Geo-Inf., 2021, doi.org/10.3390/ijgi10100668.

3. **Brishtel**, I.; Khan, A.A.; Schmidt, T.; Dingler, T.; Ishimaru, S.; Dengel, A. Mind Wandering in a Multimodal Reading Setting: Behavior Analysis & Automatic Detection Using Eye-Tracking and an EDA Sensor. Sensors 2020, 20, 2546, doi.org/10.3390/s20092546.

Conference Papers

1. **Brishtel**, I.; Krauss S.; Schmidt, T.; Rambach, J.R.; Vozniak, I.; Stricker, D. Classification of Manual Versus Autonomous Driving based on Machine Learning of Eye Movement Patterns. 2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Prague, Czech Republic, 2022, pp. 700-705, doi.org/10.1109/SMC53654.2022.9945234

Following publications belong to research activities out of the Ph.D scope :

Conference Papers

1. **Brishtel**, I.; Ishimaru, S.; Augereau, O.; Kise, K.; Dengel, A. Assessing Cognitive Workload on Printed and Electronic Media using Eye-Tracker and EDA Wristband

(2018). In Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion. Association for Computing Machinery, New York, NY, USA, Article 45, 1–2.
doi.org/10.1145/3180308.3180354

2. Loch, F.; Quint, F.; and **Brishtel; I.**. Comparing Video and Augmented Reality Assistance in Manual Assembly (2016). 12th International Conference on Intelligent Environments (IE), London, UK. doi.org/10.1109/IE.2016.31

2 Background

2.1 Forms and Functions of Mental Imagery

As motivated in Section 1, this work seeks to provide prospectives towards mental imagery-aware systems enhanced by low-cost sensors and machine learning algorithms. Considering a broad spectrum and types of mental imagery, the thesis focuses on mental imagery in the context of learning and driving activities. Before stepping forward the particular mental-imagery aware systems, this chapter provides background information about the phenomena of mental imagery and its underlying subcomponents – mind-wandering and spatial imagery. Next, sensors used in this thesis and underlying physiological mechanisms used for sensing mental imagery are introduced. Finally, state-of-the-art machine learning algorithms, performance metrics and AI-explainability methods are discussed.

2.1.1 Mind-Wandering

The phenomenon of mind-wandering is probably well-known to everybody: While reading, driving, or engaging in a routine task, people often find their attentional focus drifting to internal thoughts or concerns. Mind-wandering is defined as *"a shift in the contents of thoughts away from an ongoing task and/or external environment to self-generated thoughts and feelings"* [195]. Roughly, mind-wandering can be classified along two dimensions: *task-related thoughts* (TRTs) and *task-unrelated thoughts* (TUTs). These types affect task performance and the mental well-being of people differently. While *task-related thoughts* foster creativity and problem-solving ability [195, 216], *task-unrelated thoughts* are associated with decreasing learning performance and reading comprehension [192, 202].

As in the case of mental imagery, the quantification of mind-wandering and the

complexity of its detection are the central issues in mind-wandering research. The frequently-used technique for mind-wandering measurement is experience sampling. Experience sampling itself has various forms, including *self-caught* and *probe-caught* methods, that are the most common in the research field. See the extensive review of [228] for other methods. Both forms have their advantages and disadvantages. Thus, the self-caught method is less disruptive, allowing participants to report mind-wandering whenever it occurs freely. At the same time, it heavily depends on participants' awareness of their thoughts and therefore works for each individual differently. The probe-caught method can compensate for the lower ability to be aware of the content of own thoughts increasing the chances of revealing the overlooked episodes of mind-wandering. Nevertheless, it might be perceived as disruptive and even enhance involuntary mind-wandering [182, 228].

Research on mind-wandering plays an essential role in education since it is primarily associated with a low task and learning performance [112, 192, 194]. The attentional decoupling caused by mind-wandering is assumed to suppress information processing from external sources (i.e., learning materials), impairing learning performance. Moreover, the awareness of experiencing mind-wandering is linked to meta-awareness–the ability to track and monitor one's thought process and subjective experience [181]. This ability enormously varies among learners. On the other hand, some tasks require attentional shifting to internal thoughts for a successful task performance [194]. For example, mental arithmetic [185] as well as autobiographical memory recall all require internally directed attention. Furthermore, a few studies investigating the effect of TRTs on learning performance indicated their critical role in learning at early stages where prior knowledge is low or absent [117]. In line with that, some cognitive psychologists denoted the importance of contextual settings for the assessment of mind-wandering [78]. Unfortunately, this antagonistic property of mind-wandering is often overlooked in the context of learning systems.

As mentioned in Section 1.1, fMRI [170, 204, 201] and EEGs [216, 18, 31, 116] provide great opportunities to register the episodes of mind-wandering on the neural level. However, these methods are not transferable to daily life applications due to their complexity, size, and costs. Simpler and more scalable detection can be accomplished using eye-trackers [23, 224, 92, 190, 196, 202, 22] and the measurement of electrodermal activity [191, 24, 51]. Notably, the sensory or neuro-imagery-based

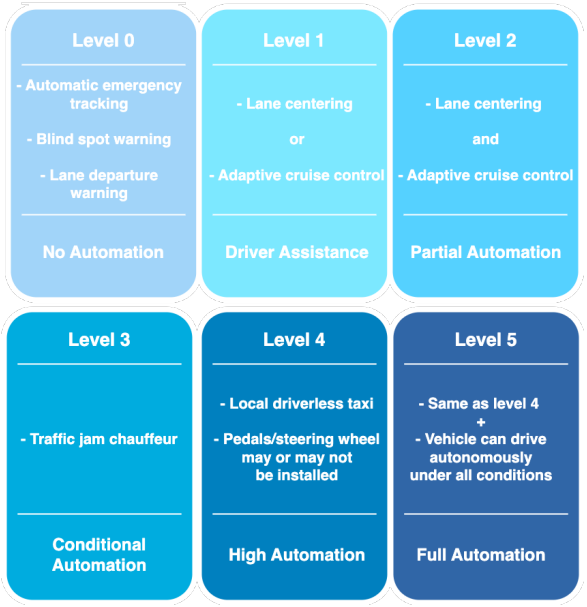


Figure 2.1: Six levels of automation und related driver assistant features. Adopted from SAE [155].

techniques still require an ES sampling technique to provide a ground truth for the experienced episodes of mind wandering.

Mind-wandering detection and intervention provide new opportunities for *mind-wandering-aware systems* that can adapt to the mental state of learners [67, 103]. As mentioned earlier, accurately detecting mind-wandering onset is still a challenging task [18, 31, 23]. However, recent studies showed the general feasibility of low-cost eye-trackers powered by machine-learning algorithms for mind-wandering detection in learning tasks [105, 67, 23, 29]. Maintaining attentional states employing mind-wandering-aware systems is supposed to support the learning process and increase overall learning performance.

2.1.2 Spatial Imagery

Navigation is one of the most crucial abilities humans have adopted from their ancestors for functioning in everyday life. Successful navigation requires contribution from multiple cognitive processes, including memory, spatial imagery and planning [26]. Spatial imagery refers to the inspection and assessment of spatial properties (e.g. distance,

relative position to any reference point) and/or the spatial manipulation (e.g. rotation, shifting, reorienting) of mentally generated images [176].

Spatial imagery retrieves from memory previously perceived spatial information and builds environmental representations. The core of spatial knowledge builds the concept of a "cognitive map" which is a mental representation of the physical environment [212]. Cognitive maps incorporate a person's spatial knowledge and thus their ability to navigate through a physical environment. The development of spatial knowledge is assumed to include three successive steps: "landmark knowledge", "route knowledge", and "survey knowledge", as proposed by Siegel and White [108, 188]. According to their framework, the development of spatial knowledge begins from landmark knowledge or knowledge about scenes or eye-catching objects within these scenes. Landmarks were found to enhance way-finding, spatial decision-making, and self-orientation in space [77, 141, 40, 225]. Next, route knowledge includes the integration of sequences of view-action pairs [125], for example "turn left to the big blue shopping centre". Finally, survey knowledge combines the observed sequences of routes and landmarks into a unified spatial reference system [148, 14], where a cognitive map emerges from.

Navigation is one of three basic visual tasks that is performed while driving [118] and takes a vital role in the hierarchical vehicle control loops [186]. Active navigation requires full engagement of the driver, including situational awareness, vehicle control, active planning and decision making, choosing and maintaining a target point in the field of view, and thus maintaining a continuous update of the visual scene [41, 10, 217]. In contrast, driving with an autopilot changes the driver's role to a passenger or a passive operator [166]. This role change results in a different task responsibility level between the system and driver, increasing passive fatigue [126] and reducing vigilance, and lowering driving engagement compared to manual driving [174]. Expanding technological innovations in the field of information technologies have enabled the widespread integration of navigation systems into mass-produced vehicles, reducing the demand for drivers to possess extensive spatial knowledge about a particular area. The primary purpose of using these systems is to provide drivers with topographical information about new or unknown locations and to improve route planning and spatial decision-making [42, 77]. In addition, modern navigation systems share information about road works, speed limits and traffic jams, which should reduce the number of traffic delays. Thus, navigation systems can reduce driver mental stress and enhance

navigation performance leading to safer driving behavior. Navigation systems are even considered ecologically beneficial, as they can significantly decrease pollution, driving time, and energy consumption [110]. While the main intention behind integrating navigation systems into serial vehicles was to mitigate drivers' cognitive and physical workload and assist them in way-finding [42, 77], significant adverse effects on spatial knowledge have also been substantiated.

The presence of a navigation system while a driving task was associated with poorer way-finding ability [151] and less precise scene recognition and ordering [42] compared to printed maps. Interestingly, navigation with a printed map was associated with the experienced increase in cognitive workload. A significantly higher cognitive workload was observed while driving with a printed map. However, this finding was inferred from post-experimental interviews and was not validated by the experimental task itself [42]. Finally, persons who travelled in a passive driving condition were observed to have sparser spatial knowledge [148]. These results suggest that frequent use of navigation systems is more likely to hinder the acquisition of spatial knowledge so that unknown environments remain unknown for a longer time [168]. Despite the large number of studies investigating the effects of navigation systems on spatial knowledge, only a few of them examined participants in real driving environments or highly immersive driving simulations [42, 9, 141, 91, 223, 96, 108].

The deployment of autonomous driving systems in serial vehicles is expected to usher in a new era of traffic navigation and driving in general. While the currently approved level of autonomy is "2" (see Figure 2.1), the vision of leading manufacturers is a fully autonomous system in production vehicles that does not expect the driver's presence in the driving loop during the entire driving time, providing the full control to the system. Despite the increasing number of studies on autonomous driving and driver behavior, their main focus has been on technology acceptance, reliance, attention [65], trust, and mental load induced by the used system [150]. In contrast, drivers' spatial cognition and knowledge have been rather neglected.

2.1.3 Spatial Imagery & Engagement

By now, no studies have been explicitly exploited the link between behavioural engagement and spatial imagery. From the studies on mental imagery (to recall the difference

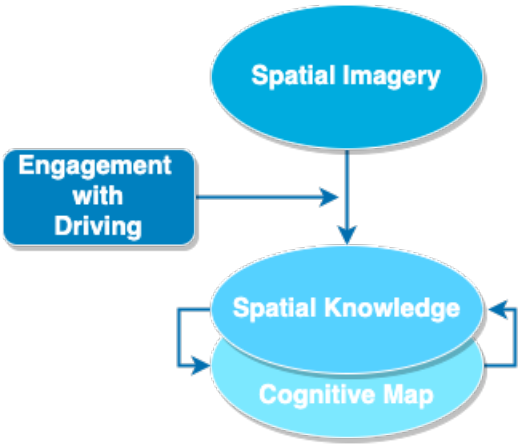


Figure 2.2: Theoretical link between engagement with driving, spatial imagery, cognitive maps and spatial knowledge proposed by the thesis. Engagement with driving moderates the strengths of spatial imagery, which activates spatial knowledge. Spatial knowledge provides a unified reference system which is integrated in the form of a cognitive map. The groove of spatial knowledge results in an updated and sophisticated cognitive map.

between mental and spatial imagery, see Section 2.1.2), it is, however, known that active dealing with a task can promote mental imagery, its vividness, and thus facilitate a planning and goal-oriented behaviour [175]. This thesis suggests that behavioural engagement in driving is a crucial component to reinforce spatial imagery and thus facilitates accessibility and further development of cognitive maps and spatial knowledge. Figure 2.2 provides a graphic representation of the proposed link between engagement, spatial imagery and spatial knowledge. Engagement itself can be easier recognized with the sensors than spatial imagery through behavioural patterns and physical activity. Thus, the driver's engagement with driving can be used to indicate active spatial imagery in a navigation task. Considering the growing demand for driver monitoring under an increasing automation level, engagement-aware systems can provide a building block to contribute to an in-car safety level.

2.2 Physiological Sensors

The following section introduces background information about the sensors deployed in this thesis to ensure the appropriateness of the selected physiological sensors. In particular, the section focuses on eye-tracking, electrodermal activity and radar technology for potential mental imagery-aware systems. It lists their relevant properties and underlying physiological mechanisms.

2.2.1 Eye-Tracking

In 1967 Yarbus demonstrated how the pattern of fixations and eye movements across the same picture depends on the viewer's intentions [233]. For instance, when the task was to judge the age of the depicted persons, observers systematically scanned their faces, whereas determining their wealth led to the scanning of clothing and furniture. This principle holds true in the context of mental imagery where gaze behavior is strongly moderated by the presence of mind-wandering [103, 190] as well as spatial imagery and engagement [211, 13, 118].

Among the variety of available eye-tracking technologies, one of the most simple and handy systems is corneal-reflection-based eye-tracking. The corneal-reflection-based eye-tracking system emits infrared light (IR) that is reflected by the cornea of the eyes and measured relative to the location of the pupil center [111, 74]. The individuals usually do not perceive the IR light; the stationary version of these eye-tracking systems does not come in direct contact with the individuals, making them comfortable for data recordings. Given an appropriate calibration of the eye-tracker, a viewer's point of regard (or the point being gazed at) can be estimated and mapped into the coordinate system of the existing system environment. Regardless of the type of eye-tracking technology, several fundamental gaze features with a strong association with the underlying cognitive functions can be extracted from nearly every eye-tracking system [235, 43, 52].

Figure 4.16 represents a gaze scan path while reading behavior. A *fixation* is an eye movement stabilizing the retina over a stationary area of interest [74]. It is also known as *gaze time* and is typically measured in milliseconds. Depending on a particular task, the threshold of a single fixation lays within a range of 200-300 milliseconds but might

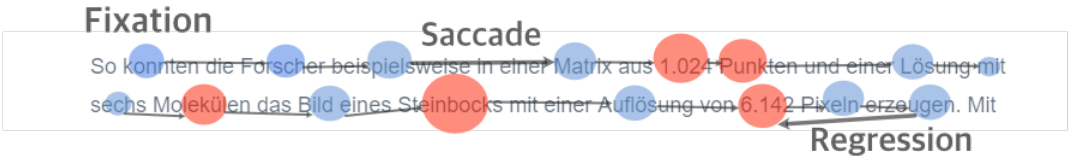


Figure 2.3: Eye movement while text reading: The size and color (red is associated with longer duration) represent fixation duration.

exceed both the upper and lower border [43]. The experiment run by Chen et al. [52] showed that the latency of *fixation duration* might be associated with the increasing load on working memory and attentional resources. They also found a positive correlation between task difficulty and fixation duration. *Fixation rate* or a *number of fixations* is the total number of fixations observed within a certain period. With the increasing difficulty of text comprehension or visual task complexity, the number of fixations was found to increase significantly [172].

A *saccade* is a rapid transition between two fixation points that takes about 30-80 ms to complete [239]. The length between two fixation points is defined as *saccade length*. Typically, for visualization of saccades, scan path graphics are used. Several studies demonstrated that *saccade duration* negatively correlates with increasing task affordance [52, 215]. Saccade length was found to be a highly discriminating indicator of cognitive workload in an experiment with variable task difficulties [52]. Saccade length also allows predicting the type of reading behavior [43]. The saccadic transition can be divided into three types: read forward, skim forward, and long skim jumps. A read forward means a normal reading behavior, whereas a skim forward or long skim jumps indicate a lack of attention to the reading task.

While reading (in European languages), any transition between two fixation points from right to left (excluding line breaks) is called *regression*. The amount and type of regressions indicate a reading or skimming behavior that might correlate with cognitive workload [43]. *Regression length* and the *number of regressions* were found to have a positive linear dependency on the mental workload. In addition, several studies reported a dependency between these two variables, lower reading comprehension and increased working memory demand.

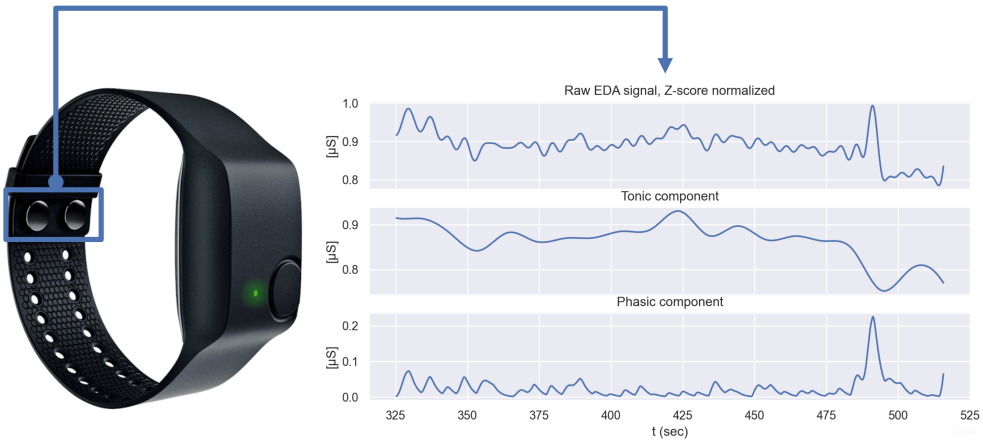


Figure 2.4: **Left**: Empatica E4 wristband with two (highlighted) snap-on silver (Ag) plated electrodes for recording of EDA signal. Image adopted from [1]. **Right**: Example of EDA signal with underlaying SCL and SCR components. (**Top**): Normalized raw EDA signal (z-score). (**Middle**): Tonic component. (**Bottom**): Phasic component.

2.2.2 Electrodermal Activity

The electrodermal activity (EDA), also known as skin conductance response (SCR) or galvanic skin response (GSR)), is an activity associated with alternations in the electrical properties of the skin mediated by the level of physiologically-induced sweating [93]. Even though sweating is primarily linked to a thermoregulation function, clinical studies investigating schizophrenia, assessment of pain, and peripheral neuropathy could state a dependency between the emotional arousal of a human and electrodermal skin response [19]. Increasing interest and broad implementation of sensors capturing EDA beyond clinical studies are due to the simplicity and non-invasiveness of data acquisition. Two electrodes placed on the skin's surface (bipolar recording) measure changes in skin resistance triggered by stimuli. Traditionally, the electrodes are placed on the thenar eminences of the palms and the volar surface of the medial or distal phalanges of the fingers [62]. Modern EDA electrodes are integrated into a wristband that can be easily placed on a participant's or patient's wrist, providing them an almost unrestricted degree of freedom in their activities (see Figure 2.4). Notably, the strength and quality of the EDA signal strongly depend on the surface where the sensor is placed: The number of sweat glands on a human wrist is lower than on the palms, resulting in a lower EDA

response.

The EDA mechanism can be described as follows: when the sudomotor nerves stimulate sweat production, the conductivity on the skin surface changes due to sweat secretion. Historically, both sympathetic and parasympathetic branches of the autonomic nervous system were assumed as potential regulators of EDA. At the end of the 20th century, studies investigating sympathetic action potentials in the peripheral nervous system while simultaneously recording EDA showed evidence for sympathetic control of EDA. Neuroimaging technology has shown several consistent patterns in the brain, where activations associated with attentional and emotional responses correlated with changes in the electrodermal conductivity [62].

A raw EDA signal consists of two main components: *Skin Conductance Level* (SCL) or tonic components and *Skin Conductance Responses* (SCRs or phasic component) [62]. SCL relates to slow drifting components, and its common properties are a gradual decrease while participants are at rest and a rapid increase when a new stimulus is presented. It is related to general information about psycho-physiological state [93] and can also be linked to the attentional state [100]. SCR is a fraction of SCL that represents small waves superimposed on the drifts in SCL and reflects sudomotor activity. The SCR is associated with short-term changes in EDA as a reaction to stimuli. The phasic component includes higher frequency components, in particular, a phasic driver. The spike density observed in SCR is linearly related to the number of recruited sweat glands and, therefore, SCR amplitude [19].

The typical size of SCL components ranges between 2 and 20 μS (microsiemens), and an increase in SCR triggered by a stimulus ranges between 0.1 and 1.0 μS. However, these values enormously vary across individuals (individual differences). Both components can be described in terms of several additional properties and subcomponents but they are not considered in this work (for more information and further reading, see: [62, 32, 93]).

2.2.3 Radar

Radar technologies experience a growing interest in human activity recognition (HAR) and human monitoring. Compared to other optical systems, radar provides unrivalled advantages in terms of privacy, robustness to environmental conditions, low sensitivity to

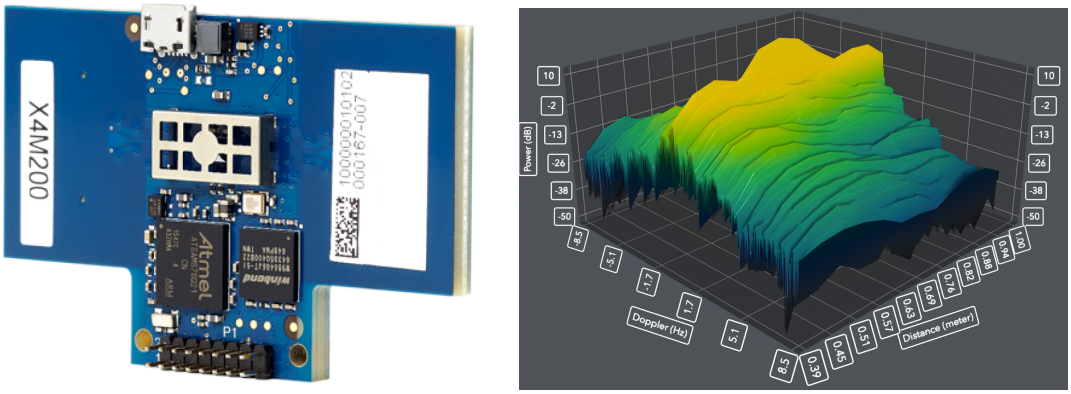


Figure 2.5: **Left**: Xethru X4M200 UWB pulse-Doppler Respiration Radar. Image adopted from [177]. **Right**: Doppler-range data recorded by Xethru X4M200 and visualized by Xethru-Explorer software.

obstacles and hazards, and usability [132], enlarging the number of potential application areas. While the central area of radar applications for HAR remains indoor activity classification [39, 30, 241, 205, 231, 183, 232, 153, 132], vital sign monitoring [145, 234, 177] and fall detection [76, 115], another prospective field is driver monitoring. Several companies have already integrated radar-based solutions for presence and seat occupancy detection [171] as well as vital sign recognition [107, 130, 147]. Recent studies also have pointed to radar systems' feasibility in recognising drivers' behavior and physical state even in moving vehicles using radar [66, 122].

A wide range of radar types can be found in the field of HAR [132]. The most common types used can be divided into two families: Continuous-Wave (CW) and Pulsed radars. Continuous-wave radars continuously transmit radio energy at high frequencies, and the radar echo is received and processed continuously. Frequency-Modulated Continuous-Wave (FMCW) radars belong to the CW group and transmit a frequency-modulated electromagnetic wave and capture its scattering from the targets. Based on the properties of the captured scattering, distance, velocity, size, and orientation of the targets can be calculated [219]. In contrast to CW radars, pulsed radars transmit briefly, followed by a long pause while the radar is in receive mode. Ultra-wideband (UWB) (see Figure 2.5) is a family of pulsed radars that transmit low-powered pulses over a wide spectrum [39]. It allows them to have a higher range resolution resulting in more fine-graded information about the target [84]. The UWB radars are able to

resolve the conflict between Doppler and range resolution while capturing the Doppler information of each scattering center of the human body [132]. Moreover, they are robust to multi-path distortion [240] and have a low energy consumption.

The research on HAR demonstrated outstanding results and multiple advantages of deep learning techniques for classifying radar data. In particular, previous studies showed that radar echo data can be treated as an image in a spectrogram or as time series of the intensity values [132].

2.3 Machine Learning for Sensory Data

In this thesis, the methods of machine learning are used to enable the classification of sensory data for the presence of episodes of mental imagery. In particular, both classical methods from machine learning and deep learning techniques are deployed. In a nutshell, machine learning is a subfield of Artificial intelligence and describes the ability of the system to acquire its knowledge by learning relevant patterns from the input data and to generalize it for new, unseen data. The classical machine learning algorithms are based on statistical learning theory, which attempts to build consistent estimators from the data [70] and requires hand-crafted features as input data calculated by engineers and researchers. In contrast, Deep Learning is a biologically motivated framework mimicking a human brain's learning ability. It can extract relevant features and patterns from raw data and use them to learn complex, high-dimensional, nonlinear mappings for a recognition task. Another characteristic of Deep learning is the gradient-based learning approach, where the network parameters can be updated with the back-propagation algorithm [129].

The selection of machine learning methods described and used in this thesis is motivated by the review of related studies in the field that are further introduced in Chapters 3, 4 and 5.1.

2.3.1 Logistic Regression

One of the most popular and straightforward machine learning algorithms is logistic regression. Logistic regression is a classification algorithm originating from statistics and used to calculate the class membership probability for one of two possible classes in the given data set using the logistic sigmoid function [70]. The logistic regression model is given by:

$$p = \frac{1}{1 + e^{-(b0+b1x)}} \tag{2.1}$$

where p is the probability of one observation belonging to one of the two classes, x is the independent variable of interest, $b0$ is the y-intercept, and $b1$ is the coefficient for the independent variable. The observed results are the ratio between the probability of an event occurring and the probability of an event against the same outcome [199]. There are several optimization algorithms used for model parameter estimation. Usually,

however, they are estimated by maximum-likelihood estimation [70]:

$$\hat{\theta}_{MLE} = \underset{\theta}{\operatorname{argmax}}, L(\theta) = \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{n} f(x_i; \theta) \tag{2.2}$$

where θ is parameter values. The main advantages of logistic regression are the low model complexity, a low number of model parameters, and thus a lower risk of model overfitting, and good model interpretability [70].

2.3.2 SVM

Support Vector Machine (SVM) was introduced in 1995 by Cortes and Vapnik for optical character recognition tasks [56]. It was based on the idea of mapping input vectors (x) in a non-linear manner to a high-dimension feature space $z = \phi(x)$ to make the separation easier. Importantly, SVM was developed for binary classification tasks. An optimal hyperplane or a decision boundary is estimated to separate the input into two classes in the resulting feature space. The SVM algorithms select that decision boundary that maximizes the margin between two classes. A margin is defined as the sum of the distances to the hyperplane from the closest points of two classes [157]. There are two types of margin: *Hard Margin SVM* and *Soft Margin SVM*, where the latter can handle noisy data and outliers using hyperparameters. However, these techniques perform well only with linearly separable data.

The non-linear mapping is performed by using a *kernel trick* – a technique that maps data from a lower dimensional space to a *characteristic space* where the data are linearly separable [179]. Typically it is performed by using the *radial basis function (RBF)* also Gaussian kernel, or *Polynominal kernel*. Thereby, RBF results in the most complex decision boundary and thus provide better classification results. RBF kernel is defined as:

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right) \tag{2.3}$$

where $||x - x'||^2$ represents the squared Euclidean distance between two feature vectors, and σ is a free kernel parameter regulating nonlinear mapping from input space [73]. After the optimal decision boundary has been estimated, it is projected on the

datasets' original space to attain a discriminant function [179]. The class prediction is performed as follows:

$$w^T\Phi(x) + b \geq \begin{cases} \geq 1 & \text{if } y_i = 1 \\ \leq -1 & \text{if } y_i = -1 \end{cases} \quad \text{i} = 1, \ldots, n \tag{2.4}$$

where $w^T\Phi(x)+b$ denotes a hyperplane consisting of feature mapping function $\Phi(x)$, weight factor w regulating orientation of the hyperplane, and a bias term b. The margin from the closest point of each class to the decision boundary is defined by $1/||w||$, and the distance between two classes is $2/||w||$.

SVM has several advantages making it incredibly popular among other machine learning techniques. Thus, it has few hyperparameters and works well when data has more dimensions than samples. Finally, due to available kernel tricks, it can handle various data types [218].

2.3.3 Random Forest

Random Forest is a tree-based ensemble learning method for classification (including multiclass classification) and regression tasks introduced by Breiman in 2001 [34]. The idea behind Random Forest is to use aggregated predictions from several randomized, uncorrelated decision trees as base learners to get the most frequently predicted class (*majority voting*) [57]. One of the main advantages of Random Forest is its robustness towards overfitting, which is achieved using two different sources of randomization: (1) training each of the individual decision trees on the independent bootstrap sample with replacement from the input data and (2) selection of the subset from predictor variables (or features) at each node split to search for the best split [21].

The major voting is performed by function defined as:

$$f(x) = \arg\max_{y \in Y} \sum_{j=1}^{J} \mathbb{I}(y = h_j(x)). \tag{2.5}$$

where $\mathbb{I}(\cdot)$ is the indicator function , $h_j(x)$ is a base learner used for building ensemble predictor $f(x)$. The margin function estimates to which extent the average number of votes of X, Y exceeds the average vote for any other class [57].

Random Forest has numerous hyper-parameters that can be tuned, making the model training process more complex than SVM 2.3.2. At the same time, using the default hyper-parameter setting for model training was shown to be associated with a solid average performance [21, 82].

The Random Forest algorithm has several considerable advantages making it for the vast number of researchers the first selection while training a machine learning model, namely: (1) relative robustness to noise in the data, (2) availability of internal estimates of error, strength, correlation, and variable importance, (3) ability to handle both continuous and categorical data, and (4) high level of parallelization [34]. Despite this method's wide range of advantages, Random Forest performs poorly when handling data with linear combinations of predictor variables [57].

2.3.4 Gradient Boosting

The Gradient Boosting algorithm was proposed by Friedman in 2001 for classification and regression tasks [86]. Boosting implies the idea of growing the trees sequentially instead parallelly. Like Random Forest, Gradient boosting uses an ensemble of decision trees for a particular learning task. However, the Gradient boosting algorithm does not rely on averaging the models in an ensemble. Instead, with each learning step, a weak, base learner model is trained regarding the error of the whole ensemble [152]. It is performed by constructing new base learners to be maximally correlated with the negative Gradient of the loss function. Thus, given a training dataset $S = \{x_i, y_i\}_1{}^N$, with input features $x = (\{x_1, \ldots, x_z\})$ and $y = (\{y_1, \ldots, y_z\})$ corresponds to the labels of the response variable, the goal is to reconstruct unknown functional dependence between the explanatory variables and the labels of corresponding response variables with the function estimate $\hat{f}(x)$:

$$\hat{f}(t) \leftarrow \hat{f}_{t-1} + p_t h(x, \theta_t) \tag{2.6}$$

such that specified loss function $L(y, f)$ is minimized [152]:

$$(p_t, \theta_t) = \arg\min_{p,\ \theta} \sum_{i=1}^{N} L(y_i, \hat{f}_{t-1}) + ph(x_i, \theta),$$

where $h(x, \theta)$ is a parameterized base learner, p_t is an iteration specified step-size.

Importantly, solving the optimization problem occurs with each p_t, which can be considered as a greedy step in finding an optimal local solution at each step. Depending on the responding variable, several families of the loss function can be used: Gaussian, Laplace, Huber, Binominal loss function, and others [152].

Unlike the bagging strategy used in Random Forest, where new models are trained separately using randomly sampled data with replacement, the boosting method can use the same training data to build new learners. Retaining training data provides a potential for overfitting: if the model is not properly regularized, the probability of overfitting increases. In addition, Gradient Boosting has more hyperparameters, including learning rate and regularization, to be tuned for a good performance. One of the main drawbacks of Gradient boosting is a high memory consumption with the increasing number of boosting iterations required for learning [152]. By properly selected hyperparameters, however, Gradient Boosting has a robust predictive performance achieved through the iteratively refined model. It can handle a mixture of data types, such as categorical, numerical, and ordinal variables providing researchers with a wide range of potential applications.

2.3.5 Convolutional Neural Network

Convolutional neural network (CNN) is one of the most commonly used neural network architectures for image classification tasks, inspired by the primary visual cortex (V1) of a mammalian brain [60], and used for processing data with a grid-like topology (including 2-D images) [90]. The first CNN architecture was introduced in 1998 by Lecun et al. [129] for a digit recognition task. The proposed architecture *LeNet* (see Figure 2.6) consisted of five hierarchical convolutional layers, followed by two

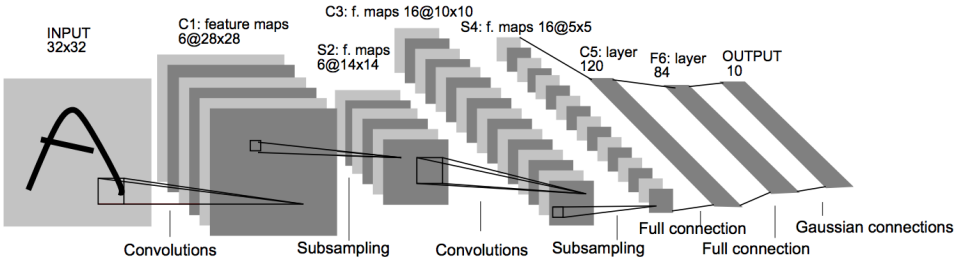


Figure 2.6: Architecture of a CNN LeNet-5. Image adopted from [129].

fully connected layers. In 2012, a new CNN architecture *AlexNet* was introduced by [124] Krizhevsky et al., demonstrating outstanding performance on the series of image classification tasks.

The convolution layer performs a convolutional operation while scanning an input X accordingly to its dimensions. The output of a convolution layer is a feature map with the equivalent dimension as X [4]. Learning local features occurs through receptive fields regulated by hyperparameters *filter size* F and *stride* S. The output of a convolutional operation is a feature map (see Figure 2.6). The pooling operation is typically applied after a series of convolutional layers to reduce the spatial size of the obtained feature map. This step minimizes the computational cost and forces the network to learn the features extracted from the previous layer. *Average pooling* and *Max pooling* are particular types of pooling where the average and maximum values are taken. Typically, max pooling is the first selection when performing a pooling operation. After a series of convolutional layers, each linear activation undergoes a nonlinear activation function, the most common activation function used for this purpose is *Rectified Linear Unit* (ReLU) [8], but also the Tanh and Sigmoid functions can be used. Finally, the feature maps are propagated to the fully connected layer (it can be one or two of them), which is a classifier, connecting the input with the output space.

2.3.6 Long Short-Term Memory Network

Long Short-Term Memory Network (LSTM) can be considered an extension of Recurrent Neural Networks (RNNs), a group of neural networks for processing sequential data $x^{(1)}..., x^{(\tau)}$. As the name suggests, RNNs have recurrent connections allowing them to

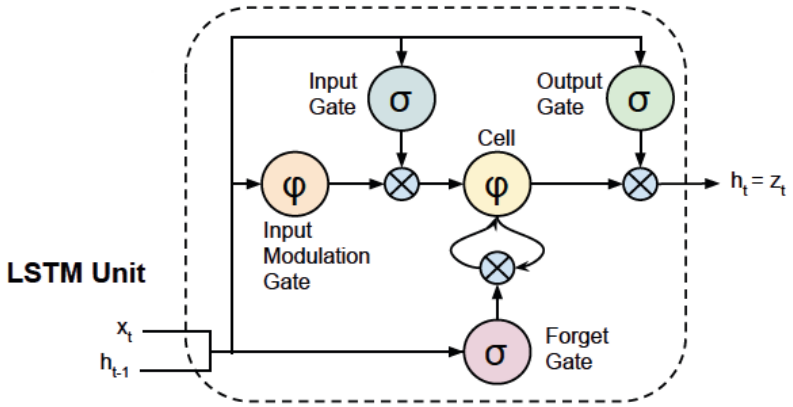


Figure 2.7: Architecture of an LSTM cell. In contrast to an RNN architecture, it has three additional gates: 1) input gate, 2) input modulated gate, 3) and forget gate. Image adopted from [68].

use their internal state, or memory, to process sequences. The LSTM was introduced in 1997 by Hochreiter and Schmidthuber [101] to overcome the error back-flow problem (vanishing and exploding gradient problem) by storing long-term dependencies. The error back-flow problem occurs when the layerwise back-propagated gradients for the weight update of the network parameters are exploding (or vanishing) due to high (or small) numbers. The storing and updating of information in the cell is achieved by deploying three gates: 1) *input gate* $i_t \in \mathbb{R}^N$, 2) *forget gate* $f_t \in \mathbb{R}^N$, and 3) *output gate* $o_t \in \mathbb{R}^N$. The forget gate estimates when the cell state should be forgotten, and the remaining two gate control input and output, respectively. Thus, the LSTM updates for a particular timestamp t follow the equations:

$$\begin{aligned} i_t &= \sigma(W_{x_i} x_t + W_{h_i} h_{t-1} + b_i) \\ f_t &= \sigma(W_{x_f} x_t + W_{h_f} h_{t-1} + b_f) \\ o_t &= \sigma(W_{x_o} x_t + W_{h_o} h_{t-1} + b_o) \end{aligned} \tag{2.7}$$

The resulting three values provide a base for calculating the internal cell state c_t and hidden cell state h_t as follows:

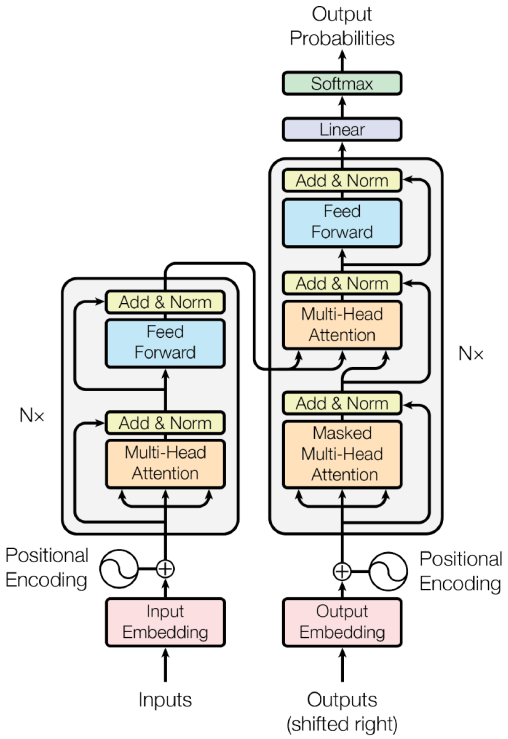


Figure 2.8: Basic Transformer-model architecture. The architecture uses stacked self-attention and fully connected layers for the encoder and decoder (left and right halves). Image adopted from [220].

$$\begin{aligned} c\prime_t &= \tanh(W_c[h_{t-1,x_t}] + b_c) \\ c_t &= f_t * c_{t-1} + i_t + c\prime_t \\ h_t &= o_t + \tanh(c_t) \end{aligned} \tag{2.8}$$

Initially proposed for sequential data, the modern modification of LSTM architectures allows coping with images and 2-D input data.

2.3.7 Transformer Network

The first Transformer network was introduced in 2017 by Vaswani et al. in machine translation [220] to resolve the need to use convolutional or recurrent layers [87]. Instead, the authors proposed the self-attention mechanism, which allows capturing global features simultaneously, resulting in a high parallelization level and increasing

performance on the data with long-range dependencies.

Figure 2.8 represents the architecture of the Transformer network. Like in sequence transduction models, the core of the Transformer builds an encoder and decoder. The Transformer encoder block extracts the features while processing the input sequence. The encoder consists of multiple identical layers comprising two sublayers: a multi-head self-attention mechanism and a position-wise fully connected feedforward network. In addition, a residual connection around each of the sublayers is used, followed by a layer normalization to attain an appropriate gradient flow and better training stability [220, 227]. The decoder block of the Transformer has an additional third sublayer which performs multi-head attention over the input from the encoder stack. The Transformer's attention is a projection of a query and a set of key-value pairs to the output [220]. The attention of the Transformer is based on *Skaled Dot-Product Attention* , which consists of three linear layers running simultaneously, namely query, key and value (Q, K, V). The attention score between a query and a key determines the importance of the key's value in computing the weighted sum. The attention function is computed parallel as follows:

$$\textbf{Attention(Q, K, V)} = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.9}$$

where Q represents the matrix with the packed set of queries, K and V are the matrices containing keys and values, respectively. d_k rerpesents a dimension. Another characteristic feature of the Transformer is a positional embedding – a dense vector incorporating a position of words in a sequence. It helps the network to learn local and global dependencies within an input sequence [87], overcoming the absence of recurrence and convolution [220].

The Transformer network proposed by Vaswani et al. became a starting point for developing the Vision Transformer (ViT), the network architecture developed for computer vision applications. Dosovitskiy et al. demonstrated that a pure Transformer applied to image patches could perform on the level of the state-of-the-art CNN networks (e.g. ResNet) in the image recognition task [69]. At the same time, it requires significantly less computational resources for model training.

2.4 Performance Metrics

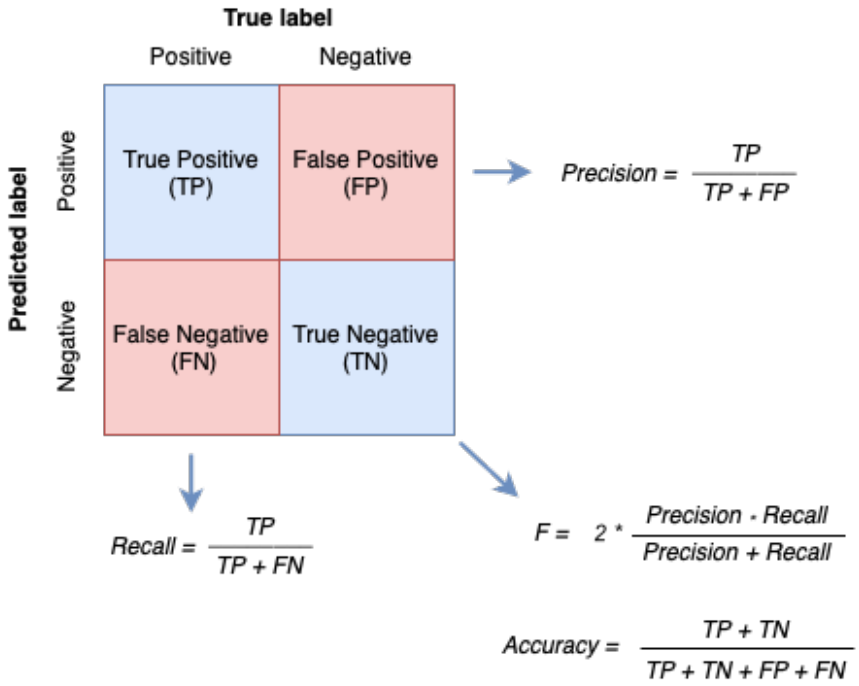


Figure 2.9: Estimation of Precision, Recall, F1score and Accuracy within a confusion matrix for binary classification problem. TP represents correctly predicted positive instances. FP represents instances that were incorrectly predicted as positive. FN are positive instances that were incorrectly predicted as negative. TN represents correctly predicted negative instances.

This thesis uses several metrics to evaluate the performance of Machine Learning algorithms deployed for mental-imagery classification tasks. The proposed metrics are selected for the classification task and based on a literature review to enhance the comparison to the state-of-the-art.

Precision and *Recall* are two frequently used performance metrics to report the fraction of retrieved data from a dataset or collection. More specifically, *Precision* refers to the fraction of true positive instances (*TP*) to all instances or labels that were predicted by the model as positive (*TP* + *FN*). *Recall* is the fraction of true positive instances to the total number of all existing positive instances in the given dataset. The value of both metrics ranges from 0.0 to 1.0, and the higher values refer to better model

performance. These metrics are primarily for binary classification tasks and undesirable for unbalanced datasets with uneven class distribution.

F_1-score is a harmonic mean of the Precision and Recall. It can be used to report binary and multiclass classification task results. It provides a more balanced summarization of the classification performance and therefore is frequently deployed in cases with imbalanced datasets [95]. In this case, *weighted*-averaged F1 score can be used, where metrics for each class are calculated and then averaged by the number of true instances for each class.

Yet another frequently used metric for evaluating classification models is *Accuracy*. It measures the proportion of true instances retrieved, both positive (TP) and negative (TN), among all data from the dataset [58]. Accuracy is a reasonable choice when the classes in the dataset are balanced and have equal importance. The value range of Accuracy is between 0 and 1, where 1 means the highest Accuracy. The calculation of F1-score, Precision, Recall, and Accuracy are summarized in Figure 2.9.

Area Under the curve (AUC) describes the capability of the classifier to distinguish between classes and is a summary of *Receiver Operating Characteristic* curve (ROC curve). ROC shows how the number of correctly predicted positive examples varies with that of incorrectly predicted negative examples [61]. AUC generates a probability curve with sensitivity (True positive rate) versus Specificity (True Negative Rate) within the value range of 0 and 1.

A *confusion matrix* is used to evaluate the consistency of a classification model with the ground truth data while estimating its accuracy [99]. It has a size of *m* x *m*, where m represents the number of defined classes in the dataset. The confusion matrix provides a detailed breakdown of the model's classification performance and can be used for binary and multiclass classification models. Figure 2.9 represents a confusion matrix for a binary classification task.

2.5 Model Explainability

The complexity of the machine learning algorithms frequently prevents researchers and engineers from interpreting the outcomes obtained from a particular model. However, modern research standards expect machine learning models to achieve a high prediction

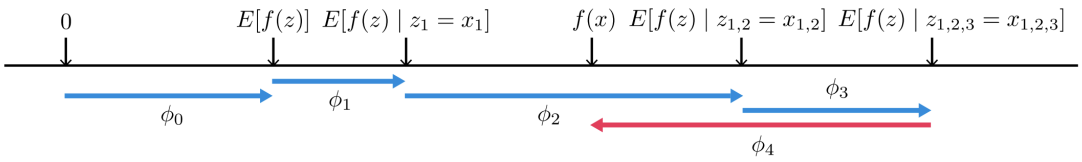


Figure 2.10: SHAP values estimation schematic overview. Image adopted from [140].

power and generalization level and be highly interpretable [137]. For the interpretation of the deployed machine learning models from Sections 3 and sections 4, this thesis focuses on the *SHAP values* method.

Introduced by Lundberg and Lee [140], the SHAP values are an additive explanation approach based on the *Cooperative game theory* (Shapley values) and propose a unified method for calculating feature importance. This method uses the contribution of the features within a given dataset to the model output to explain the observed prediction [142]. More precisely, the SHAP method approximates Shapley values [184, 142] since their exact computation is a challenging task because of the exponential complexity. The approximations through SHAP are performed through several techniques, including special weighted linear regressions or different assumptions about feature dependence for ensemble tree models [142].

The sum of the SHAP values of each feature is equal to the final prediction. In this case, a SHAP value is not the difference between the prediction with and without a feature but a feature's contribution to the difference between the actual prediction and the mean prediction.

The estimation of SHAP values is performed by solving the equation:

$$
\begin{aligned}
f(h(x(z'))) = E[f(z \mid z_S)] &\text{ SHAP explanation model simplified input mapping} \\
= E_z\left[\mathbb{E}[f(z)] \mid z_S\right] &\text{ expectation over } z_S \text{ given } z \\
\approx E_z\left[\mathbb{E}[f(z)]\right] &\text{ assume feature independence} \\
\approx f\left([z_S; \mathbb{E}[z_S]]\right) &\text{ assume model linearity}
\end{aligned}
\tag{2.10}
$$

where z_S represents missing values for features not in the set S.

Figure 2.10 represents a diagram for calculating SHAP values. SHAP values attribute to each feature value explaining how to get from the base value $E[f(z)]$ that would be predicted if none of the features to the current output $f(x)$ is known. ϕ_i is the impact of including or excluding feature i on the model's performance. Thus, the outcome of the Shapley value can be interpreted as a contribution ϕ_i of the value i compared to the average prediction for the given dataset. The Shapley values work for classification and regression tasks making their application area broad.

3 Towards Mind Wandering - Aware System

3.1 Proposed Study

Mind wandering detection and intervention provide new opportunities for the learning assistant systems where the goal is to adapt the learning material and information representations to the mental state of learners. Accurate detection of mind wandering onset is a challenging task [18, 31, 23], but recent studies showed the general feasibility of mind wandering detection in learning tasks by using low-cost eye-trackers powered by machine learning algorithms [105, 67, 23, 29]. Maintaining attentional states employing context-aware technologies and adaptable presentation systems is supposed to support the learning process. In some situations, mind wandering may be desirable and necessary for successful learning. A few studies investigating the effect of TRTs on learning performance indicated their critical role in learning at early stages where prior knowledge is low or absent [117]. In line with that, cognitive psychologists such as Faber et al. [78] identified the importance of contextual settings for assessing mind wandering.

This chapter introduces low-cost and scalable techniques for deploying a mind-wandering-aware system. Thereby, reading is selected as a context since it is essential to any learning process. This chapter consists of two parts. The first part introduces a behavioral analysis and focuses on the effects of semantics and music on the frequency and content of mind wandering in a reading task. This step is essential to evaluate the user behavior in a given context properly, thus increasing the study's reliability level. The second part addresses the feasibility of sensory and behavioral data combined with machine learning models to classify mind wandering in the learning context. The

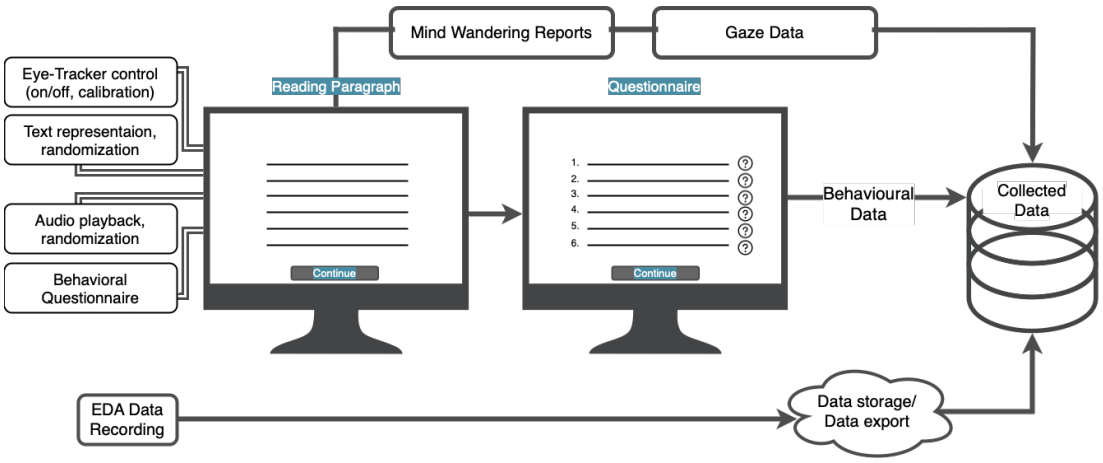


Figure 3.1: Overview of the proposed system for the collection of the episodes of mind wandering in a learning scenario. Beside the acquisition of mind wandering using self-reported method, the system controls audio playbacks and text representation and collects behavioral data.

content introduced in this chapter is published in Brishtel et al. [35].

3.2 Related Work

Triggers

The episodes of mind wandering were found to be induced by multiple internal and external stimuli. Thus, there is strong evidence that interest [88], task difficulty, emotional state [195, 204], monotone [194], and distractibility [85, 229] are strong determinants of mind wandering. Episodes of mind wandering can also occur in response to a semantic stimulus if the latter evokes memories [17, 78]. Interestingly, the temporal focus of mind wandering was found to be influenced by the experience with a topic [193]. Thus, people with a corresponding background experience episodes of mind wandering more frequently. Faber and D'Mello investigated relationships between stimulus type and the content of mind wandering [78]. Their experiment with 88 participants demonstrated that the content of mind wandering is widely spread across multiple thought categories and associated with various triggers. Semantically rich content was found to force mind wandering associated with memory retrieval.

Along with semantics, music was also found to impact the frequency and content of mind wandering [81, 204, 47]. Taruffi et al. [204] showed that music triggering sad, low-arousal emotions caused more frequent episodes of mind wandering by their participants, compared with music triggering happy, high-arousal emotions. Feng and Bidelman's study [81] investigated the relationship between music and mind wandering in the context of a lexical processing task. Their results yielded similar findings as the study of [204], showing that mind wandering occurred more frequently in conditions with unfamiliar music. The authors suggested that the increased frequency of mind wandering resulted from boredom, negative mood, and distractibility as a response to lower emotional arousal in unfamiliar music. The nature and content of mind wandering is, therefore, complex and multifaceted, and treating this mental state as merely detrimental to the performance can be misleading. There needs to be more information about possible interaction effects between different contextual factors on mind wandering.

Eye-Tracking & Mind Wandering

A strong relationship exists between the episodes of mind wandering and eye movement patterns [22, 224, 92, 190, 196, 214, 202]. In particular, fixations, blink rate, pupil diameter, eye vergence, and saccades were found to be strong indicators for the presence of mind wandering. Benedek et al., for instance, provided a comparative analysis of oculomotor behavior in conditions with internally (IDC) and externally (EDC) directed cognition using anagram and sentence generation tasks [20]. The results showed that IDC was associated with fewer and longer fixations, higher variability in pupil size diameter and eye vergence, and a lower angle of eye vergence. In a meditation task with a self-caught sampling method, Grandchamp et al. found a significantly smaller pupil diameter during episodes of mind wandering [92]. Smilek et al. investigated changes in a blink rate during mind wandering in a reading task using an auditory probe-caught method [196]. They found a higher blink frequency during mind wandering than when participants were focused on a task. The abovementioned changes in oculomotor behavior may indicate attentional decoupling and the resulting suppression of the visual input within the episodes of mind wandering (for more details, see [181]).

Using neuropsychology findings and machine-learning techniques, several research

groups succeeded in building automated eye-based detectors of mind wandering. For example, Bixler and D'Mello [23] recruited 178 participants and asked them to read four texts and report mind wandering using the self-caught method. Among other features, they used 40 global gaze features to build a user-independent mind-wandering detector. The reported models achieved accuracy between 67% and 72%. Unfortunately, the authors used a high-end, expensive eye tracker, which restricts the deployment of their method on a large scale.

Addressing this issue, Hutt et al. [104] demonstrated the feasibility of using low-cost eye trackers for automatic mind wandering detection in a classroom setting. Using eye tracking data of 135 high school students recorded during computerized learning, their model achieved an F_1-Score of 0.59.

Electrodermal acitivity & Mind Wandering

So far, only a small number of studies used EDA as a neural marker for mind-wandering detection. The first studies employing EDA for mind wandering detection pointed at induced alternations in dermal properties during episodes of mind wandering. Thus, Blanchard et al. used EDA, skin temperature (ST), and context features (i.e., text, timing, and text difficulty) to build a supervised classification model for automatic mind wandering detection in a learning setting [24]. Using the *Affectiva Q* with a sampling rate of 8 Hz, they recorded data from 70 undergraduate students using a combination of self-caught and probe-caught methods. From EDA and ST, authors calculated the following physiological features: the standardized signal, an approximation of the time-derivative of the signal, the frequency and magnitude from Fast Fourier transformation, the spectral density of the signal and the autocorrelation of the signal at lag 10 (for a comprehensive review see [24]). Then, for the physiological features, the mean, standard deviation, maximum, ratio of maxima, and the ratio of minima were calculated, resulting in 43 features for EDA and ST, respectively. Context features included 11 elements. The model using the combination of features from EDA and contextual features achieved the highest kappa coefficient of 0.15. Kappa ranges from 0 (chance agreement) to 1 (perfect agreement). Following a classification by Landis and Koch [127], a value of 0.14 indicates "slight agreement". The combination of EDA, ST, and context features achieved a "moderate agreement" with a kappa of 0.22. Unfortunately,

their work did not consider a solely EDA-based classification model.

Cheetham and colleagues [51] investigated the feasibility of using only EDA features for the automatic detection of mind wandering in the context of meditation. Using peaks observed from a low-pass filtered EDA signal as a feature, the authors reported an area under the curve (AUC) of 0.81. AUC ranges from 0.5 (chance) to 1 (perfect agreement). An AUC of 0.81 is better than Blanchard et al.'s kappa of 0.22, but still needs to be higher. However, the episodes of mind wandering and resulting physiological changes during meditation might differ from those in tasks requiring higher cognitive functions [204]. Nevertheless, both of these studies did not consider the underlying sub-components of EDA in their analysis.

3.3 System Design

For the text, music, comprehension questions, and a behavioral questionnaire representation, eye-tracker control, and collection of mind wandering reports Electron based [75] (see Figure 3.1). It allowed exact control over the experimental flow (including text and audio randomization) and ensured the proper data recording, including eye-tracking data. The collected data were accordingly stored for further user-model evaluation and the building of a machine learning model. EDA data were recorded and stored separately since this system design of the sensor does not provide direct access to the data storage and thus could not be integrated into the developed system. Self-caught reports included timestamps of the episodes of mind wandering, behavioral data included responses from a questionnaire relating to text and self-perception. The exact experimental flow is described below.

Apparatus The reading task was performed on an NEC MultiSync EA241WM monitor with a resolution of 1920x1200 pixels operating at 60 Hz. The distance between the participant and the monitor was fixed at 60 cm (see Figure 3.2). A Tobii 4C gaming eye-tracker [2] with 90 Hz sampling frequency and a scientific license was used to capture eye movement. A head chin rest was used to achieve high accuracy of the eye-tracker and avoid any head movement artefacts. To measure the EDA, an Empatica E4 wristband [1] with a sampling frequency of 4 Hz was used. The wristband E4 does

not require any calibration. It was placed on the wrist of the non-dominant hand. Since the fingers have the highest density of sweat glands, lead wires were used for the index and ring fingers to acquire a higher EDA signal. The lead wires were snapped instead of plated electrodes. The data recording continued throughout the experimental session.

Research Design

To investigate the effects of the context factors on the frequency of mind wandering, two independent variables were defined, namely *Text Type* and *Music Type*, and their levels were additionally manipulated. Three different music types were differentiated: Sad, Happy and No-Music as a control condition. The factor Text Type contained three levels: Computer Science, Psychology and Random Text Type. Six dependent variables, Interest, Difficulty, Tiredness, Perceived Mood, Attentional Focus and Type of Thoughts, were acquired from behavioral data and are discussed below. A repeated-measure design where each participant underwent all experimental conditions were used. Thus, the proposed experiment has a 3×3 repeated-measures design (see Table 3.1). On each level, self-report, behavioral data and physiological measures were used to assess the ongoing mental state.

Table 3.1: Experimental Design: experimental and control conditions with a total sample size.

3 × 3		**Music Type**		
		Sad	Happy	No-Music
Text Type	Psychology	80	80	160
	Computer Science	80	80	160
	Random Topic	80	80	160

Participants

21 graduate and undergraduate students (17 male) with an average age of 25.3 ($SD = 2.6$) years were recruited from the University of Kaiserslautern at the Department of Computer Science via a mailing list. All participants were native German speakers with normal or corrected-to-normal vision. All students had a major in computer science. For

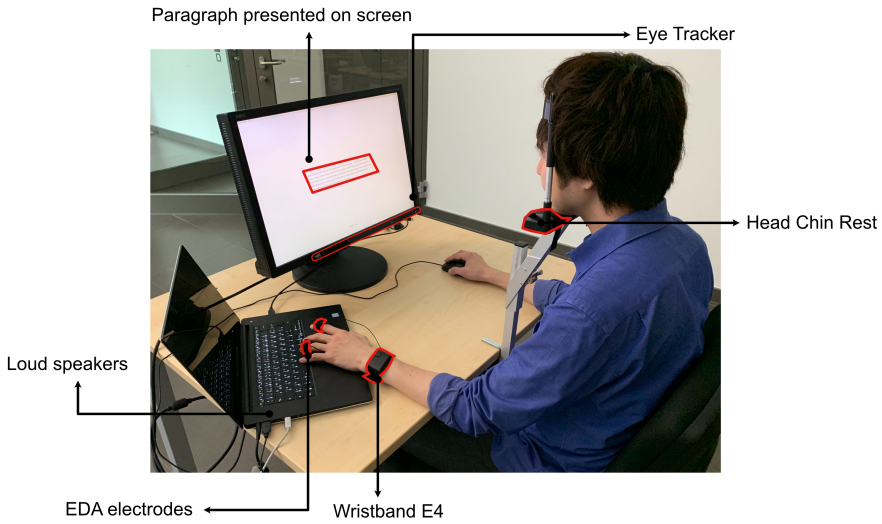


Figure 3.2: Participant performing the reading task. Eye movements were collected using an eye-tracker Tobii 4C, EDA was recorded with a wristband Empatica E4. Occurrences of mind wandering were collected through self-reports. Image adopted from [35].

participation in the study, participants were offered either course credits or a 10-euro gift card.

Reading Material

12 scientific texts in the German language were selected from an online platform [230] that publishes articles in popular science. Since one of the goals of this study was to investigate the role of semantic information on the frequency and content of mind wandering, eight texts from the categories of Psychology and Computer Science were deliberately selected that were assumed to match the personal and academic background of our participants. Texts in Computer Science included topics about Storage and Algorithms, Cryptography, and Artificial Intelligence. Texts in Psychology included topics about Names and Stereotypes, Conscious and Unconscious Self, and Self Perception. The remaining four texts belonged to the category of Random Text Type and consisted of topics with a low probability of matching one's personal or/and academic background. These texts included information about Photosynthesis, Metaphors, Bovine Diseases, and the Australian Football game "Footy". All the texts had a comparable

length ($Mean = 230.4, SD = 19.6$) and difficulty level (LIX score [167] $Mean = 57.3, SD = 9.3$). The average text difficulty resembled the university literature for home works or exams. In the next step, each text was split into four paragraphs of comparable length (mean number of words = 60.9, SD = 6.3). The effectiveness of this text-splitting strategy for mind-wandering research was demonstrated in related studies [22, 24]. Thus, one experimental session contained two texts from each text type, split into four paragraphs presented one by one. This resulted in 24 reading segments (2 text x 3 categories x 4 paragraphs). The order of text types was randomized within each participant. The text paragraphs were presented in the center of the computer screen in a Sans-Serif font, font size 18 pixels, and in black color.

Music Stimuli

Audio stimuli were selected from the pool provided by Taruffi et al. [204]. Their stimulus pool was pretested for homogeneity of happy and sad compositions with significant differences to the opposite affective tone. The set of their sad stimuli was rated as highly pleasant, slightly arousing, clearly sad, and unfamiliar. For the happy conditions, the musical stimuli were rated as highly pleasant, very arousing, clearly happy, not sad, and unfamiliar. In this study, unfamiliar musical stimuli (without vocals) were selected to avoid possible memory effects on mind wandering. Our experimental stimulus set included eight happy and eight sad compositions. Each of them was clipped into 45-second segments and normalized by following the documentation provided by *Audacity* [16]. In the next step, sad and happy music segments were randomized among paragraphs with respect to their condition and counterbalanced with respect to the No-Music conditions. Thus, the participants read 12 paragraphs with and 12 paragraphs without music in the background in one session. The onset and offset of the music playback was synchronized with each text paragraph's beginning and end. The music volume was set to a comfortable level and was kept at the same level in all conditions. Laptop speakers were used to play the music (see Figure 3.2).

Procedure

The experiment was performed in accordance with the institutional ethical guidlines of the German Research Centre for Artificial Intelligence. Before starting the experimental

session, participants were provided general information about the purpose of the study. After they gave their informed consent, the experiment was started. Participants were also informed that they could withdraw from the study at any point in time. The experimental room was quiet, and light, temperature and possible noise distractors within it were controlled.

The experiment was performed in accordance with the institutional ethical guidelines of the German Research Centre for Artificial Intelligence. Before starting the experimental session, participants were provided with general information about the purpose of the study. After they gave their informed consent, the experiment was started. Participants were also informed that they could withdraw from the study at any time. The experimental room was quiet, and light, temperature and possible noise distractors were controlled.

The experiment was split into two 45-minute sessions. Each session was run on different days to avoid fatigue and interaction effects between Happy and Sad conditions. Depending on the Music Type playing in the background, each session belonged to the Happy or Sad condition. Prior to the experimental session, participants received the following instruction:

> *You will be presented with six different texts split into paragraphs. Each text contains 4 paragraphs. Please read each paragraph as attentively as possible. Ignore possible music in the background. While reading text, your attention might drift from reading to internal thoughts or concerns, which is totally natural. If it happens, please press the space button and focus back on the reading task.*

Next, participants were asked to take a comfortable sitting position which they could keep for at least the next ten minutes. To prevent any discomfort that could restrict normal reading behavior, the chin rest was adjusted for each participant individually. To avoid motion induced artifacts in EDA, participants were asked to keep their hand with the attached wristband on the table. After each text (or four paragraphs), participants were allowed to make a short break, where they were reminded to report mind wandering in the reading task. The eye-tracker was re-calibrated after each break using a standard 5-point method. To ensure attentive reading behavior of participants, there was a comprehension question after each paragraph with one possible correct

Table 3.2: Questionaire used for collection of behavioral data including text relevance, text perception and direction of experienced thoughts.

Question	**Scale**
Q1. How interesting did you find the last paragraph?	0%: not interesting at all, 100%: very interesting
Q2. How difficult did you find the last paragraph?	0%: not difficult at all, 100%: very difficult
Q3. How tired did you feel while reading the last paragraph?	0%: not tired at all, 100%: very tired
Q4. What was your level of happiness while reading the last paragraph?	0%: not happy at all, 100%: very happy,
Q5. What was your level of sadness? while reading the last paragraph?	0%: not sad at all, % 100%: very sad
Q6. Did the context of the last paragraph match your academical or personal background?	0%: no matched at all, 100%: fully matched
Q7. While reading the last paragraph, where was your attention focused?	-5: fully lost in thoughts +5: fully focused on the text
Q8. While reading the last paragraph, did you have some text-related thoughts.	yes/no
Q9. While reading the last paragraph, did you have some text irrelevant thoughts (i.e., personal worries, future planning, dreams, thoughts about your relatives or friends)?	yes/no

answer, “yes” or “no”. In the next step, the participants rated each paragraph on Interest, Difficulty, Tiredness, Personal/Academic Relevance, Perceived Mood (Sadness and Happiness), Attentional Focus and Type of Thoughts (task-relevant or task-irrelevant) (the questionnaire was adopted with changes from Giambra and Grodsky [88]) (see Table 3.2).

For the questions *Q1 – Q6*, a 5-point rating scale with step sizes of 20% was used. Question 7 *Attentional Focus* was used as an additional retrospective sampling method [228] to increase the awareness about the content of own thoughts. If participants rated

their attentional focus with less than +3 (step size 1 point), two additional questions related to the type of thoughts (*Q8* and *Q9*) were displayed:

The proposed questionnaire is considered further as *behavioral data*. Additionally, time spent to read one paragraph was recorded within each participant and considered as a behavioral feature *Reading Duration*. The questions *Q4* and *Q5* are not in the scope of the analysis in this study.

3.4 Statistical Analysis

To ensure the correctness of the assumptions regarding the personal and external triggers of mind wandering by the users, this section provides results of statistical analysis of behavioural data. To examine possible influence of text and music type on the frequency and content of mind wandering, within-subject contrast effects were tested using a two-way ANOVA (analysis of variance) for a two-factorial repeated-measure design. One participant was discarded from the final analysis after he admitted to having misunderstood the experimental instructions. The final sample size used for the statistical analysis included 20 participants.

The mean values of each variable underwent a planned pairwise comparison as outlined in Table 3.1. All reported F-values for the main and contrast effects are Greenhouse-Geisser corrected. All significant tests are reported. The simple effects within each factor were analyzed using Helmert contrasts where the first two non-control conditions (Happy and Sad Music, Computer Science and Psychology Text Types) were compared, followed by the comparison of those two conditions with the control (No-Music, Random Text Type).

Personal/Academic Relevance

In order to confirm the validity of assumptions for personal/academic text relevance (PAR), the effect of Text Type on perceived PAR was examined. An ANOVA yielded a significant main effect of Text Type on PAR (see Table 3.3). The contrast analysis showed that the paragraphs from Computer Science (*mean* = 52.08, *SD* = 29.08) were rated as significantly more relevant than those from Psychology (*mean* = 20.98, *SD* = 25.29). The relevance of the paragraphs from Random Text Type (*mean* = 6.17, *SD*

Table 3.3: Analysis of Variance. Bold font denotes main and interaction effects.

Source	df	F	p
PAR			
Text Type	1.9	79.7**	0.01
C vs P	1.0	64.3**	0.01
RT vs (C+P)	1.0	94.3**	0.01
MW			
Music Type	1.6	3.5*	0.05
S vs H	1.0	0.4	0.52
(S + H) vs NM	1.0	10.3**	0.01
Text Type	1.7	1.2	0.30
Text x Music	2.7	2.8*	0.05
S CS vs (S RT + S P)	1.0	6.6*	0.02
H RT vs (H CS + H P)	1.0	5.1*	0.04
TRTs			
Text Type	1.7	4.3*	0.03
R vs (C + P)	1.0	6.1*	0.02
Music Type	1.7	4.3	0.11
Text x Music	2.8	1.2	0.31
TUTs			
Text Type	1.9	0.1	0.87
Music Type	1.4	0.6	0.48
Text x Music	3.1	0.8	0.53

PAR: Personal/Academic Relevance. MW: Mind Wandering. NM: No-Music, S: Sad, H: Happy. C: Computer Science. P: Psychology. RT: Random Topic. * $p < 0.05$, ** $p < 0.01$.

= 13.74) was perceived as significantly lower than those from Computer Science and Psychology (see Table 3.3).

Table 3.4: Average frequency of mind wandering and TRTs by experimental condition.

Condition	Mind Wandering M(SD)	TRTs M(SD)
Sad Computer Science	0.34(0.38)	0.77(0.88)
Sad Psychology	0.18(0.26)	0.87(0.98)
Sad Random Topic	0.19(0.27)	0.91(1.05)
Happy Computer Science	0.22(0.35)	0.47(0.67)
Happy Psychology	0.25(0.42)	0.53(0.76)
Happy Random Topic	0.39(0.52)	1.01(1.06)
No music Computer Science	0.12(0.18)	0.60(0.72)
No music Psychology	0.08(0.11)	0.60(0.75)
No music Random Topic	0.15(0.14)	0.87(0.92)

M: mean; *SD*: standard deviation.

Frequency of Mind Wandering

Table 3.4 represents the average mind wandering frequency reported within a particular condition. A significant main effect of Music Type on mind wandering was observed (see Table 3.3). Although there was no significant contrast effect between Happy and Sad Music, participants experienced episodes of mind-wandering significantly more frequently in conditions with music than in conditions without music (see Table 3.3). There were no significant effects of Text Type on the mind-wandering frequency. Nevertheless, there was a significant interaction effect between Music Type and Text Type on mind wandering. Thus, while reading texts in Computer Science and listening to Sad Music, participants reported more episodes of mind wandering compared to the conditions with the same Music Type but texts in Psychology or Random Topic. For the Random Text Type, this effect was the opposite: Participants reported mind-wandering significantly more frequently while listening to Happy Music compared to Sad Music and the control condition.

Task-Related Thoughts/ Task-Unrelated Thoughts

To meet ANOVA requirements, the binary variables *Q8* and *Q9* were transformed, using arcsine transformation:

$$Y' = \begin{cases} 2arcsin(\frac{1}{2N}), & \text{if } y = 0 \\ 2arcsin(1 - \frac{1}{2N}), & \text{if } y = 1 \\ 2arcsin(\sqrt{Y}), & \text{otherwise} \end{cases}$$

where *Y* denotes an observed response, and *N* the total number of all observations.

There was a significant main effect of Text Type on TRTs. The contrast analysis showed that reading paragraphs from Random Topics was associated with a significantly higher number of TRTs (see Table 3.4) compared to the residual text types (see Table 3.3). The main effect of Music Type on TRTs was not significant, no significant interaction was observed. For TUTs neither significant main effects nor interactions were observed.

Correlation Analysis

A multivariate Pearson's R analysis was run to investigate possible correlations among behavioral variables (see section 3.3) and mind wandering. Figure 3.3 represents the obtained correlation plot. Importantly, significant correlations within behavioral data are not further discussed here as they are not in the scope of this work. Only significant correlations with the Pearson's r coefficient $\geq \pm$ 0.2 are considered here.

There was a highly significant, moderate and negative correlation between mind wandering and Attentional Focus. Attentional Focus had a statistically significant but low to moderate positive correlation with Personal/Academic Relevance. A significant, moderate, positive correlation was observed between Attentional Focus and Interest. Moderate to low negative correlations were observed between Attentional Focus, Tiredness, and Difficulty. There were highly significant, moderate to low, negative correlations between Attention Focus, TRTs, and TUTs. A statistically significant, moderate, negative correlation between TRTs and Interest was observed. Significant, small, positive correlations between TRTs and Tiredness and Difficulty were present. Finally, a positive, moderate correlation was present between the frequency of mind

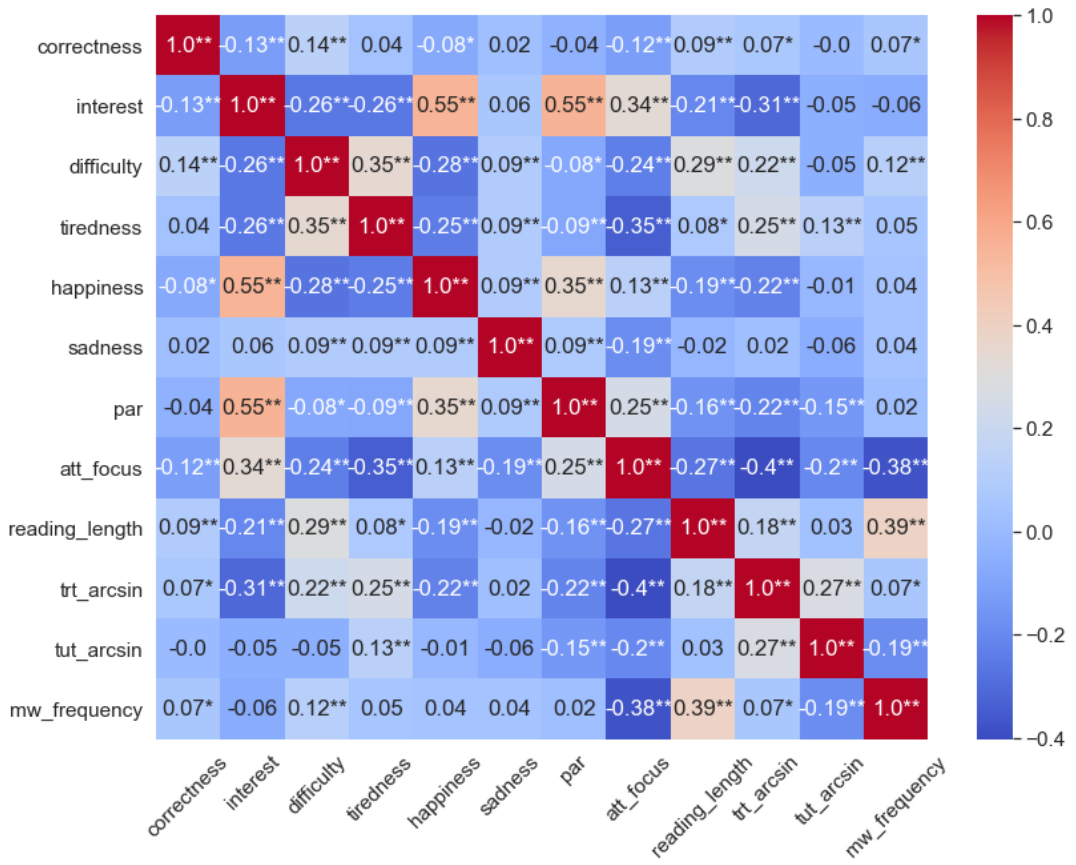


Figure 3.3: Pearson's R correlation analysis for behavioural data and reported episodes of mind wandering. **: significant on $\leq 1\%$ level; *: significant on $\leq 5\%$ level.

wandering and the reading length of a paragraph.

3.5 Machine Learning Approach

Before building machine learning models to classify episodes of mind wandering, several preprocessing steps on sensory and behavioral data were carried out. Figure 3.4 represents the pipeline for sensory data synchronization, data preprocessing, and feature engineering. All steps were performed using Python's libraries *numpy*, *pandas*, *scipy*, *sklearn* and *matplotlib*. Thus, in the first step, all recorded data were synchronized using the Unix timestamp. The sensory and behavioural data related to the paragraph reading were extracted in the second step. Next, signal and data preprocessing were

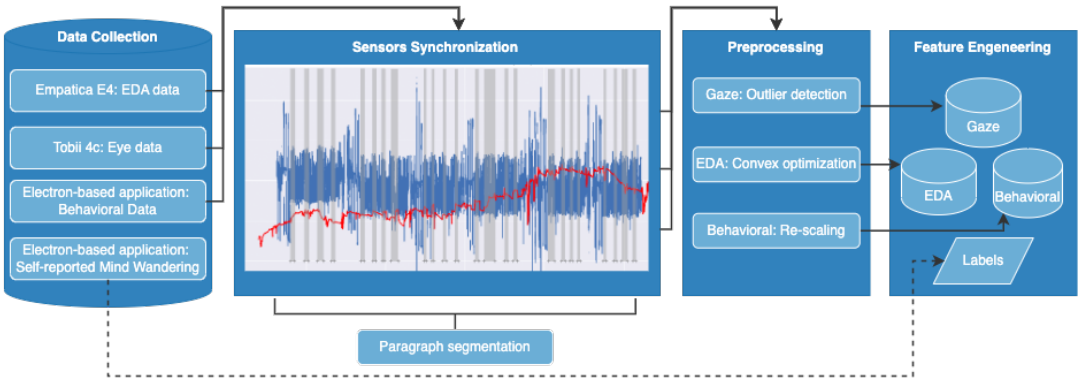


Figure 3.4: Data collection, synchronization and preprocessing pipeline. Self-reported episodes of mind wandering were used as ground truth.

performed in accordance with a signal type. Finally, the processed data were used for feature engineering. The sections below introduce a detailed overview of preprocessing steps and feature engineering for the particular data type.

3.5.1 Feature Engineering

Eye-Tracker Feature

Tobii Pro software was used to extract the raw eye data and information about the pupil diameter. Clustering of the raw gaze points into fixations was performed using the Dispersion-Threshold Identification algorithm [180]. In the next step, fixations points lying outside the reading area were removed. 59 paragraphs with insufficient recording accuracy of gaze data were excluded from the classification tasks. The final data set size contained eye movement and relating behaviour data from 871 paragraphs. Figure 3.5 represents visualized eye movements within the paragraph with reported mind wandering versus normal reading behavior.

For each paragraph, fixation-related features such as the mean fixation points etc. were extracted (see Table 3.5). In the next step, saccades and saccade-related features, such as mean saccade length, number of regression, etc., were calculated on the paragraph and participant level. Table 3.5 provides a detailed description of each included feature. In total, 19 eye movement features were calculated.

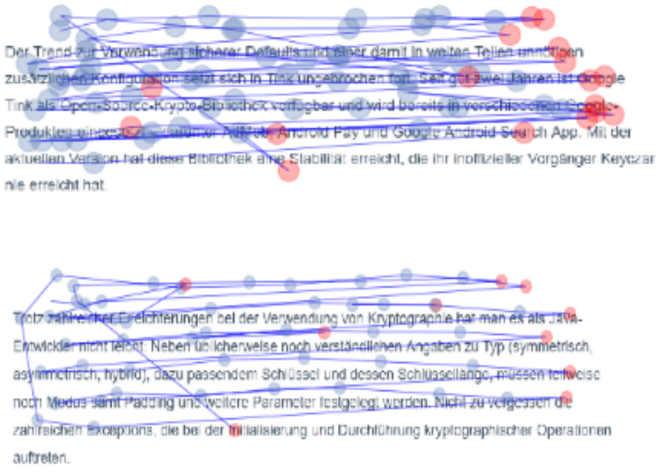


Figure 3.5: Eye movements during the reading task (fixation points in blue and regression points in red). **Top**: Paragraph with reported mind wandering. **Bottom**: Paragraph with focused reading behaviour. Image adopted from [35].

Table 3.5: Eye movement feature description. For all features mean was calculated, for bolded features min, max values were additionally calculated.

Feature	**Description**
Fixation duration	Duration of a fixation point in msec
Pupil size	Diameter of pupil in pixels (z-score)
Saccade length	Distance in pixels between 2 subsequent fixations
Saccade velocity	Transition between 2 subsequent fixations in msec
Saccade angle	Angle between x axis and the ray to the point (x, y)
Regression length	Backward transition between 2 fixation points in px
Number of regressions	Total number of regressions within 1 paragraph
Number of fixations	Total number of fixation points within 1 paragraph
Number of saccades	Total number of saccades within 1 paragraph

EDA Feature

As mentioned in Section 2.2.2, the EDA signal consists of two main components: *Skin Conductance Level (SCL)* and *Skin Conductance Responses (SCRs)*. To extract these components from the raw EDA signal, the Convex Optimization approach was used. The convex optimization approach proposed by Greco and colleagues [93] is a common technique to extract underlying tonic and phasic components. This method postulates that EDA can be described as a function of three components: a slow tonic,

the output of convolution between an infinite impulse response function (IIR), and a sparse non-negative sudomotor nerve activity (SNMA).

In line with the recommendations on signal preprocessing prior to convex optimization [93] for EDA decomposition, a z-standardization was applied to the acquired signal on the individual level In the next step, statistical features from extracted EDA components were calculated. Table 3.6 provides an overview of calculated features for raw signal, tonic and sparse components. The generated data set included 18 features in total. For the sparse component, peak amplitude (A), number of peaks within one paragraph (N_A), and the number of peaks above 1µS ($N_{A>1}$µS) [45] were additionally calculated. The 59 paragraphs excluded because of the low eye-tracking accuracy were also excluded from EDA model building. Figure 2.4 represents the results of the Convex optimization method applied to a raw EDA signal.

Table 3.6: EDA extracted features.

Component	**Feature**
Raw z-standardized	$\overline{X}$, σ, min, max
Tonic	$\overline{X}$, σ, min, max
Sparse	$\overline{X}$, σ, A_{min}, A_{max}, N_A, $N_{A>1}$µS

3.5.2 Model Building

The user-independent classification of the episodes of mind wandering while reading paragraphs where achieved by deploying machine learning models. Depending on participants' responses, each reading paragraph received either "1" if mind wandering was reported there otherwise "0" (see Figure 3.5). The data of one participant were excluded since he had not reported any episodes of mind wandering. The final data set used for model training and evaluation contained data from 19 participants.

Three machine learning algorithms were selected to classify episodes of mind wandering: Logistic Regression (baseline classifier), Support Vector Machine (SVM), and Random Forest. Before model building, training and testing data were separately normalized using *StandardScaler* to achieve equal scaling among input features. Next, because the class "1" was reported in 15.8% of all paragraphs only, after splitting

all data into training and testing sets, the training dataset was oversampled using the synthetic minority over-sampling technique (SMOTE) [50]. The same approach for the particular classification was deployed by Bixler and D'Mello [23].

To prevent model overfitting and to ensure that the best hyper-parameters are selected to build the models, a nested cross-validation procedure [48] was performed. The nested cross-validation procedure is required to avoid model biasing, resulting in overly-optimistic scores that cannot be handled by single cross-validation (not-nested). The outer loop in nested cross-validation is for model assessment, and the inner loop is for hyperparameter tuning.

3.5.3 Baseline Classification

Table 3.7: Baseline classification results for logistic regression and feature sets. Bold font represents the best classification $F1 - Score$ performance. The Fusion of Eye, EDA and Behaviour features achieved the highest classification accuracy.

Model	Feature	Kappa	Accuracy	AUC	F_1-Score
Logistic Regression	Eye	0.25(0.21)	0.72(0.14)	0.70(0.12)	0.75(0.12)
	EDA	0.23(0.28)	0.70(0.17)	0.66(0.19)	0.73(0.16)
	Fusion$_{Eye,EDA}$	0.26(0.22)	0.73(0.13)	0.71(0.16)	0.76(0.11)
	Fusion$_{Eye,EDA,Behavior}$	**0.31(0.27)**	**0.76(0.13)**	**0.72(0.17)**	**0.79(0.10)**

Logistic regression was selected as a baseline classifier for classifying the episodes of mind wandering. The data preprocessing steps before the model building were equivalent to those described in Section 3.5.2. Since logistic regression has no essential hyperparameters to be tuned, the main focus was ensuring testing data is withheld from training by deploying the leave-one-participant-out cross-validation method. The validation was repeated 19 times following the number of participants.

Table 3.7 represents baseline classification results using single gaze and EDA data sets, their fusion, and the fusion of physiological and behavioral data. Because of an unequal distribution of classes (84.2 % of all data belonged to the paragraphs without reported mind wandering) and to enhance the comparison to the state-of-the-art [24, 23], four evaluation measures are provided: F_1-Score, Receiver Operating Characteristic curve (AUC), Accuracy and Kappa. For better readability, all results are

Table 3.8: Classification results for SVM and Random Forest. Bold font represents the best classification performance. The fusion of Eye and EDA Features achieved the highest classification accuracy. Random Forest models demonstrated the highest F_1-Scores.

Model	Feature	Kappa	Accuracy	AUC	F_1-Score
Random Forest	Eye	0.25(0.21)	0.80(0.09)	0.66(0.12)	**0.80(0.09)**
	EDA	0.15(0.15)	0.83(0.08)	0.62(0.13)	**0.78(0.08)**
	Fusion$_{Eye,EDA}$	0.29(0.27)	0.83(0.08)	0.65(0.16)	**0.83(0.08)**
	Fusion$_{Eye,EDA,Behavior}$	0.31(0.27)	0.76(0.13)	0.69(0.15)	**0.82(0.09)**
SVM	Eye	0.26(0.24)	0.78(0.13)	0.68(0.13)	0.78(0.11)
	EDA	0.26(0.23)	0.73(0.14)	0.67(0.16)	0.76(0.12)
	Fusion$_{Eye,EDA}$	0.37(0.27)	0.79(0.15)	0.73(0.17)	0.80(0.12)
	Fusion$_{Eye,EDA,Behavior}$	0.41(0.28)	0.80(0.14)	0.77(0.14)	0.82(0.07)

further reported using F_1-Score. Thus, the highest observed $F1 - Score$ of 0.79 was achieved through the fusion of eye, EDA and behavioral data. Probably, the observed correlation among several input features (see Figure 3.3) could explain this moderate classification performance, as Logistic Regression does not perform optimally when high correlation structures underlay the input features [179].

3.5.4 Results

Table 3.8 represents the classification performance of the two machine learning models, Random forest and SVM, for automatic classification of mind wandering. The lowest classification accuracy was observed by SVM (see table 3.8). Unlike Logistic Regression, SVM is not sensitive to the correlation between features by selecting an appropriate kernel and performs better with heterogeneous data [179]. Considering the fusion of sensory and behavioral data (*Eye + EDA + Behavior*), Random Forest achieved the highest F_1-Score of 0.83. Notably, compared to SVM and Logistic Regression, Random Forest consists of a significantly larger number of hyperparameters resulting in a higher computational cost when tuning them [200]. Interestingly, using the combination of gaze, EDA and behavior data (*Eye + EDA + Behavior*), Random Forest was outperformed by SVM (see Table 3.8 Kappa, Accuracy and AUC). The fusion of sensory and behaviour data in SVM and Random Forest models outperformed the defined baseline by three percentage points.

3.5.5 Model Explanability

As mentioned in Section 2.5, the classification performance itself does not contribute to the model interpretability, particularly in cases where a large set of heterogeneous data is used. To understand how individual features contribute to the classification performance, *SHAP values (Shapley Additive Explanations)* [140] method was deployed. The feature importance was evaluated for each Random Forest model type tuned with the best hyperparameters.

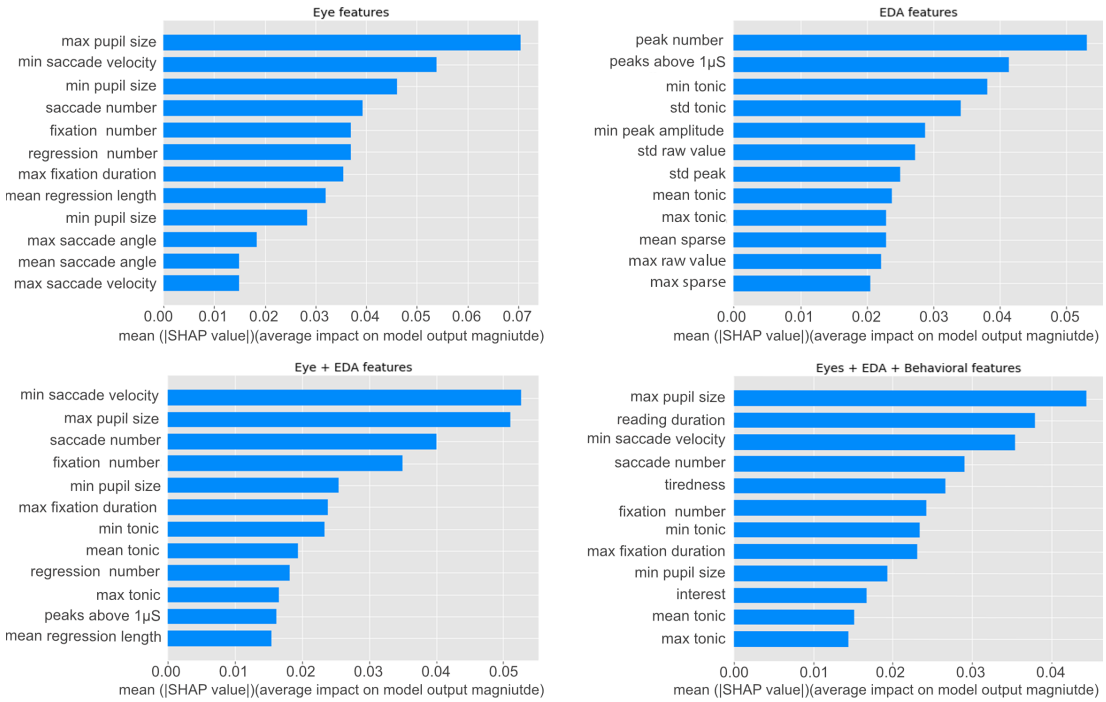


Figure 3.6: Feature importance graph for the Random Forest classification models using the SHAP method. **Top left**: Eye-based model. **Top right**: EDA-based model. **Bottom left**: Eye and EDA-based model. **Bottom right**: Sensory and Behavior-based model.

Figure 3.6 represents the SHAP values for the four models depending on the used features. The eye-based model achieved an F_1-Score of 0.80 (see Table 3.8). *Pupil size*, *saccade velocity*, *number of saccades* and *number of fixations* had the highest importance for the ability of the model to discriminate the episodes of mind wandering from the normal reading behaviour (see Figure 3.6). The observed feature importance resem-

bles the finds of prior studies [92, 20, 214]: Statistical features of pupil size, saccades and fixations were observed to be sensitive to the occurrence of mind wandering.

Next, the EDA-based model achieved a slightly lower F_1-Score of 0.78. The *number of peaks* had the highest contribution for the prediction of mind-wandering followed by the number of *peaks above 1 μS*, and the *min value* of the *tonic component*. The latter two features showed comparable feature importance. The combination of EDA and eye features resulted in an improvement of the classification performance by five percentage points achieving the highest observed F_1-Score of 0.83 among all models (see Table 3.8). The data from both sensors could bear complementary information to each other, capturing different dimensions of the participant's mental state, thus enhancing the overall classification performance. Thus, electrodermal activity can be moderated by emotional arousal triggered by the episodes of mind wandering, whereas eye movements are related to readers' visual attention and reading comprehension. Another critical point is a difference in the speed of physiological processes: while changes in eye movements caused by cognitive processes are on a millisecond level (and even on a microsecond level) [74], whereas for EDA it might take three seconds and more before any stimuli-driven alternations in the signal can be differentiated from the basic tonic activity [62]. Finally, for the Random Forest, fusing behavioral data with the sensory as proposed earlier ($Eye + EDA + Behavioral$) did not improve classification performance, albeit reading duration was the second important feature for classifying mind wandering. In contrast, it provided a countereffect where the F_1-Score dwindled at one percentage point (see Figure 3.6).

3.6 Conclusion

This work demonstrates the feasibility of EDA sensor data and eye movements in combination with machine learning for user-independent mind-wandering quantification in the context of a mind-wandering aware system. Electrodermal activity is related to emotional arousal triggered by episodes of mind wandering, whereas eye movements are related to the visual attention of learners. The model based on these two feature sets achieved the highest classification accuracy for mind wandering and outperformed the eye-based model by three percentage points. At first glance, this improvement can

appear relatively moderate. It points, however, to the general possibilities for combining features acquired from an eye-tracker and an EDA sensor. Moreover, combining the two sensors can be helpful because they operate in different time frames: the changes in eye movements caused by cognitive processes can be observed on a millisecond level [74], whereas the alternations in EDA can be visible up to three seconds after a stimulus occurred [62]. Therefore, it is presumed that when reading longer texts, the combination of EDA and eye-tracker may result in higher classification accuracy because the shortness of the paragraphs may have led us to miss some information in the EDA if the mind-wandering occurred towards the end of the paragraph. The further inclusion of behavioral features did not show a remarkable contribution to detecting mind wandering, albeit reading duration strongly contributed to the classification accuracy. This result shows that for an automatic mind-wandering detector, the sensory data are sufficient, eliminating the necessity to use additional behavioral questionnaires and thus reducing the amount of user-related information for a potential mind-wandering aware system.

The study presented here has several limitations. First, the used self-caught method cannot guarantee that all experienced episodes of mind wandering were reported. This method's reliability depends entirely on participants' awareness of the context of their thoughts. Second, the proposed method is an offline mind-wandering detection technique. A moment-to-moment detection of mind wandering using EDA features is a challenging task since, as mentioned earlier, the physiological time course of electrodermal responses varies from individual to individual and has a considerable delay between the onset of mind wandering and related alternations in the signal. Finally, the proposed multimodal sensor setup for mind wandering detection was investigated in a controlled experimental setup. Verifying these finds in a "wild" setting will be necessary. In addition, this study used a binary classification to detect the presence or absence of mind wandering. Future research must investigate the possibilities to identify the particular type of mind wandering (TRT or TUT). This distinction is important from a pedagogical point of view, where task-unrelated thoughts are a part of metacognition about unfamiliar, demanding, or tedious tasks and may contribute to the ultimate learning success. This aspect should be considered while building mind-wandering-aware systems.

4 Towards Engagement - Aware System

4.1 Proposed Study

The extended use of navigation aids and autonomous driving systems may come at the expense of reduced navigational competence in the driver. Highly automated navigation aids and autopilots can be expected to reduce mental workload in the short run, but they may be detrimental to fundamental spatial skills in the long term, for instance, in way-finding and homing behavior, route planning, distance estimation, and the general development of cognitive map of the current environment [209, 42, 168]. Such deficits could be of great practical importance when drivers have to switch to a non-automated vehicle when an automated device fails or when a detour becomes necessary that the automated system is not yet prepared for.

This chapter seeks a way to build an engagement-aware system. For this purpose, in the first step, a user model under autonomous versus manual driving mode is investigated to understand how different driving modes impact the spatial imagery of drivers. Without these insights, reliable deployment of spatial imagery or engagement imagery-aware systems is not possible. For this purpose, an extensive study with a mounted, highly immersive driving simulator and a virtual unfamiliar environment is performed, where passive transportation (autopilot) is contrasted with active driving assisted by a navigation system and a printed map. Active driving is associated with high engagement levels, and passive navigation is associated with low engagement levels. Next, based on the outcomes of the user-model study, an automized, user-independent engagement-aware system is proposed, where the particular driving mode level is classified using eye-movement data. The main results of this chapter have been published

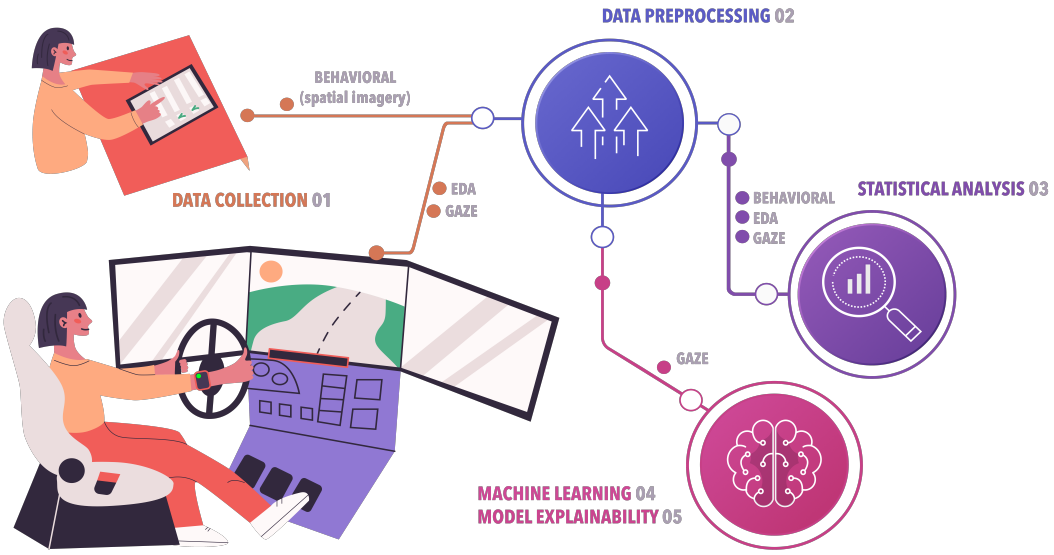


Figure 4.1: Overview of the successive steps for system design: Collection of sensory (gaze and EDA) and behavioral data, data preprocessing, statistical analysis, machine learning and model explainability. Image partially adopted from: Adobe Stock.

in Brishtel et al. [38] and Brishtel et al. [37].

4.2 Related Work

Under increasing automation levels, the driver-aware systems attempt to ensure the driver's presence in the driving loop. For this purpose, several technologies, including system messages [149], hands-on-wheels recognition and in-cabin driver monitoring camera [207, 206, 222] were integrated into production cars. Each technology has its strengths, but the considerable weaknesses are associated with perceptual failures [149], privacy issues or lack of information regarding the driver's ongoing attentional and engagement level with driving [221]. For the latter one, eye-tracking technology can be a good solution. Taking into account that failures in visual attention are responsible for a significant number of traffic accidents [118], there is no wonder that this sensor is widely employed in autonomous driving research and partly integrated into serial vehicles [149, 128, 77]. For instance, system messages might harm drivers' attention,

including perceptual failures or inappropriate [149]. Studies showed that two seconds of off-road glances are already enough for an increased risk of accidents [135, 221]. Despite the high precision of distraction recognition, in-cabin cameras are considered by many drivers as an intrusion into their privacy. While meeting the privacy requirements, the hand-on-wheel recognition technology does not provide reliable information about the driver's attentional state and engagement with the driving task [221]. In laboratory settings, it was shown that 'hand-on-wheel' had not been a reliable predictor for collision avoidance on the road.

Spatial knowledge & Navigation Modes

Qin and Karimi [168] proposed one of the first studies that explicitly investigated the differences between cognitive maps gained while autonomous and conventional (manual) driving. In their work, the experience of driving in autonomous mode was treated as an equivalent form as those when driving as a passenger. The online survey with 204 participants run in their study showed that individuals who often travel as passengers have trouble navigating using their cognitive maps. Additionally, they are more prone to get lost in a direct neighbourhood. The authors also suggested a possible reduction in spatial knowledge for drivers switching to autonomous driving.

Related studies are consistent with the assumption that the driver's role in automotive driving will change to "passenger-like". Using a survey and a map drawing task with 101 London residents, Minaei examined the effects of different travel modes on cognitive map development [146]. The author found a positive correlation between using a car in active mode and the number of roads recalled in the map drawing task. Mondschein et al observed comparable outcomes. [148]. They contrasted active car drivers with adults who used public transport more frequently than a car. It was shown that car drivers were significantly more accurate in estimating driving distance than public transportation users. It was also demonstrated that active car drivers were more likely to identify their home position using a landmark. Stülpnagel and Steffens found that backseat drivers of a tandem bicycle in a passive navigation condition had better landmark recognition than an active navigation condition [223]. Similarly, travelling by bus through a city is presumed to allow passengers to observe more of the environment and to attend more to the landmarks [96]. Bus drivers were observed to develop better

survey knowledge of a city compared to their passengers, who could gain only route knowledge [54, 15]. Unfortunately, with the exception of Stülpnagel and Steffens [223], none of the aforementioned studies deployed a real or a simulated driving scenario. Instead, they evaluated surveys or compared the spatial knowledge of active drivers and passengers, who, in principle, have a different perception of the environment. Even taking into account the view that in autonomous driving, the driver will change his or her role to that of a passenger, an examination of the possible changes in spatial learning under autonomous driving conditions in a more realistic situation is crucial.

Spatial imagery & Electrodermal activity

Navigation systems are developed to support drivers in way-finding and spatial decision-making tasks, which in turn should reduce drivers' mental and physical workload [42, 94, 160]. Similar effects are expected through the utilization of highly or fully autonomous driving systems. Not surprisingly, studies in human factors and ergonomics observed a reduction in the mental and physical workload of the human operator while using autonomous systems [158, 7]. Surprisingly little is known about cognitive workload in the context of navigation and spatial learning. Evans et al. observed that stress can be beneficial for the formation of cognitive maps [77]. Based on this observation, Burnett and Lee hypothesized that the initial demand for cognitive map development using printed maps is high [42]. With growing spatial knowledge, this demand significantly decreases.

One of the most established techniques for measuring cognitive workload is the NASA-TLX test that captures a self-reported level of the experienced cognitive demand required to solve a particular task [98, 97]. Along with this self-report method, physiological and neurological signatures can be used to indicate increasing workload and stress levels. Together with the heart rate, EDA is one of the most frequently employed physiological properties to track the driver's stress level [150, 53, 138].

Eye-tracking & Spatial imagery

Liang et al. investigated how the engagement level in not-driving-related tasks (NDRT) prior to a takeover request (TOR) influences takeover performance and situational awareness [134]. They asked 30 participants to accomplish three driving tasks on SAE

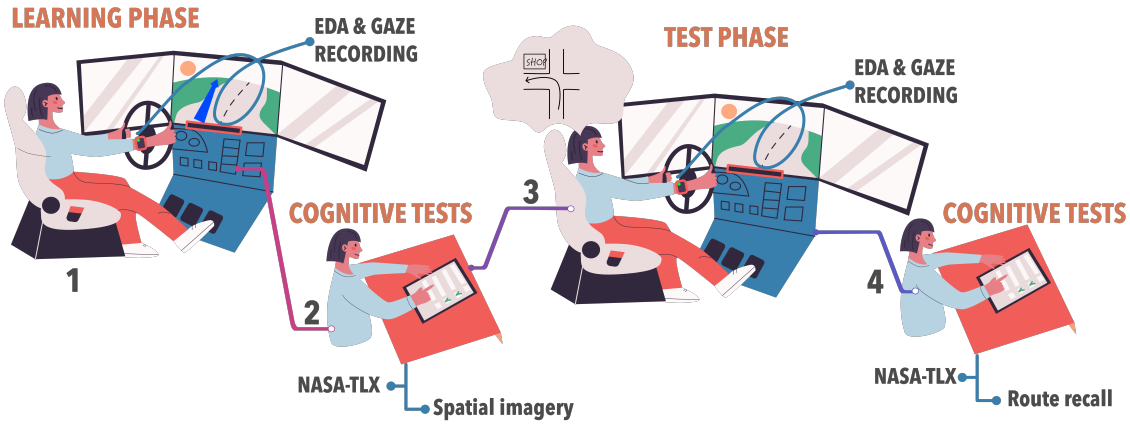


Figure 4.2: Schematic representation of experimental steps in learning and test phases. Image partially adopted from: Adobe Stock.

Level 3. In each of the driving tasks, they had to engage in three NDRT tasks prior to TORs: 1) Surrogate reference task (searching for a target circle among distractor circles); 2) Peripheral detection task (responding to visual stimuli presented in peripheral vision); and 3) a monitoring task. Among other measures, authors used pupil size, fixation duration, fixation number, a fraction of time in the area of interest (AOI), scan path length, gaze count in AOI, and spatial density to access drivers' states before and right after TOR. The gaze was mainly off-road in the surrogate reference task (SRT). In contrast, in a monitoring task, the driver's gaze distribution covered more ambient space and dispersed more than in SRT. The authors concluded that a longer time for scene viewing and more dispersed attention allocation are associated with better situational awareness. Louw and Merat observed similar results [139]. They investigated horizontal and vertical gaze dispersion in conventional and automated (SAE Level 2) driving. The authors found that during manual driving, the gaze is less dispersed.

4.3 Research Design

The experiment consisted of a learning phase and a test phase. In the learning phase, participants sat in a driving simulator. They were either passively driven by autopilot through a simulated town or drove the car themselves using either a navigation system or a printed map. In the test phase, they were required to repeat the route from the

Table 4.1: Experimental design: experimental conditions with the total sample size.

2 × 3		**Phase**	
		Learning	Test
Assistant Type	Autopilot	20	20
	Navigation System	20	20
	Map	20	20

learning phase by driving the car themselves without any navigation aids. The learning phase also consisted of a memory test requiring them to identify and order landmarks along the route.

To examine the impact of different assistant systems on the drivers' spatial knowledge and cognitive workload, two independent variables were varied: assistant type (autopilot, navigation system, or printed map) and phase (learning or test) (see Table 4.1). To characterize the behavioral performance in the test phase, five dependent variables were employed: proportion of correctly selected landmarks, or "hits"; proportion of selected foils, or "false alarms"; proportion of landmarks placed in the correct order; and number of wrong turns. Furthermore, the NASA-TLX score was measured during both the learning and test phases to trace cognitive workload. While driving tasks, gaze data were continuously recorded using in OpenDs integrated software. In addition to the behavioral and eye-tracking data, skin conductance, was recorded during both study phases and decomposed into two components, the tonic component (SCL) and phasic component (SCR). These components were derived from the EDA signal as a neurological marker of the induced stress level. In addition, participants were asked to draw a map of the virtual city, but this dependent measure is not the focus of the current paper because they are not related to the hypotheses described above. A repeated-measures within-subjects design was used where each participant underwent all experimental conditions. In such designs, statistical power depends jointly on the number of participants and the number of trials per participant and condition [197]. Figure 4.1 provides an overview of the system and data recording pipeline.

Participants

A total of 22 participants were recruited in the German Research Center for Artificial Intelligence (DFKI) from different departments via an employee mailing list. Recruitment occurred under the restrictions of the COVID-19 pandemic, which avoided bringing unnecessary visitors into the DFKI facilities. Inclusion criteria were the absence of motion sickness, a minimum of two years of driving experience, and a driving license valid in Germany. Despite being informed about the inclusion criteria, two participants reported severe symptoms of motion sickness. The experiment was then immediately stopped, and their recorded data were discarded from further analysis. The residual 20 participants (7 female) were between 22 and 50 years of age (*mean* = 29.9, *SD* = 6.5) and drove a car on average at least five times a week. All participants had normal or corrected-to-normal vision. Participants were not informed about the goals of the study prior to the experimental session. Participation was voluntary and compensated through working hours.

Driving Simulator

Figure 4.3 shows a schematic representation of the technical setup of the driving simulator. The mounted driving simulator was assembled from a Jaguar XJ 4.2 V8 Executive cockpit and an integrated input controller set (Logitech G27 Driving Force), which includes a steering wheel and two pedals for throttle and brake but no clutch pedal. The transmission was set to automatic mode so that no gear shift was necessary and the simulator could be operated without extensive training. A projector with a resolution of 1920×1080 (Full HD) displayed the driving scene on a large screen about two meters in front of the participant. The experimenter's place was set behind the driving simulator. The driving simulator setup resembled those of Feld et al. [80]. The virtual city consisted of three maps representing different parts of the city. Each map included buildings, traffic lights, road signs, road markers, and pavements with bus stops, trees, and small parkways but did not include any pedestrians or traffic.

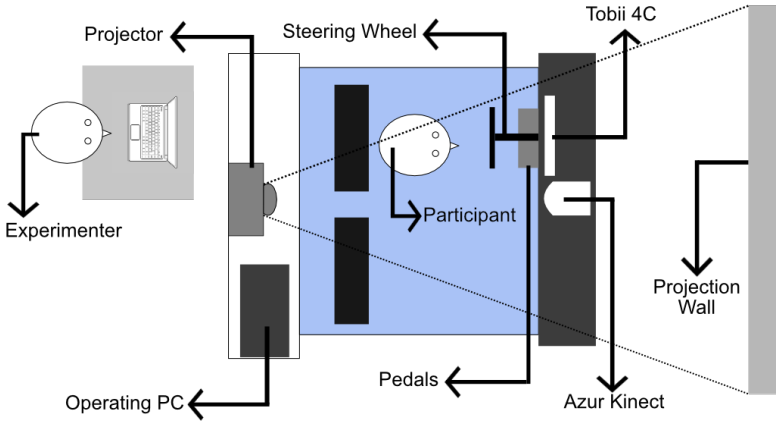


Figure 4.3: Schematic representation of the technical setup in the driving simulator.

4.4 System Design

Sensors

A wrist-worn Empatica E4 device was used to trace the participant's electrodermal activity during the driving tasks. Besides sensors for skin conductance, the Empatica E4 wristband has three more integrated sensors: a photoplethysmography sensor, a 3-axis accelerometer, and an optical thermometer. By default, Empatica E4 transmits the data from all four sensors. The EDA signal is conducted through two snap-on silver (Ag) plated electrodes with a sampling frequency of 4 Hz. The wristband does not require any calibration procedure. It was placed on the volar surface of the wrist according to the manufacturer's recommendations after the training phase in the driving simulator. The data recording continued over the entire experimental period.

Along with EDA, eye movements were recorded using a Tobii 4C gaming eye-tracker with a default sampling frequency of 90 Hz under a scientific license. The OpenDS software supports gaming eye trackers, including Tobii 4C. It provides processed eye-tracking data, including fixations, saccades (a transition between two successive fixations), and information about the gazed objects in the virtual environment. To achieve the highest calibration and recording accuracy, the eye-tracker was individually set for each participant and re-calibrated every time before the start of a driving task.

Figure 4.4: Three printed maps corresponding to the driving routes employed. To compensate for the relative shortness of the second map (**top right**), the change of the driving direction (**blue highlight**) was integrated into the driving route.

Virtual Environment and Driving Assistants

The open-source simulation software OpenDS Pro Complete [156] was used to simulate the driving environment. This application provides high flexibility for creating and customizing controlled driving environments, including map creation and 3D model integration. OpenDS is based on jMonkeyEngine V3.2 [113] with a variety of supported features, open standards (openDrive, openSCENARIO, SCENIC), and supported hardware (VR glasses, motion platforms, input controllers, etc.). The system also supports gaming eye-trackers (e.g., Tobii X and Tobii 4C) and provides processed eye-tracking data, including fixations, saccades, and information related to the fixated objects in the virtual environment.

The city center of the German city of Saarbrücken was taken as a prototype for the virtual city. From the city plan, three maps with three driving routes of comparable

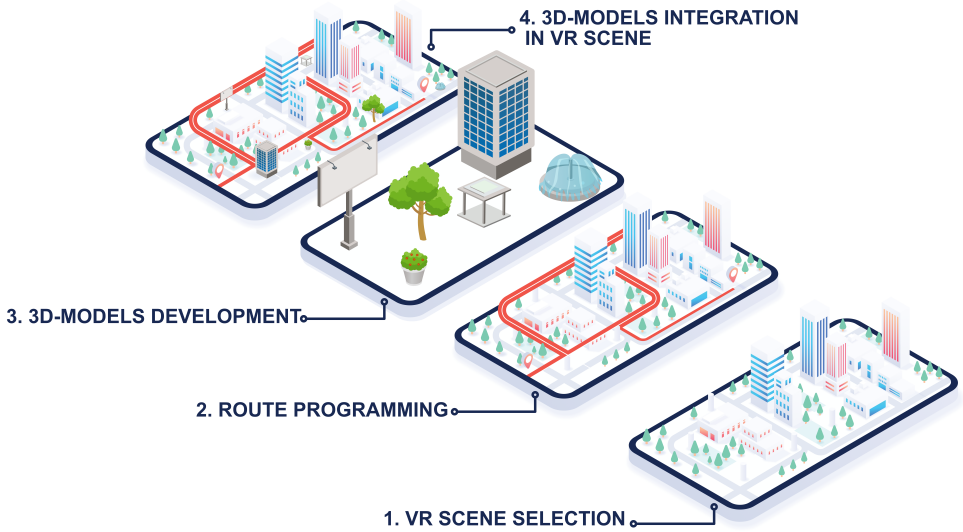


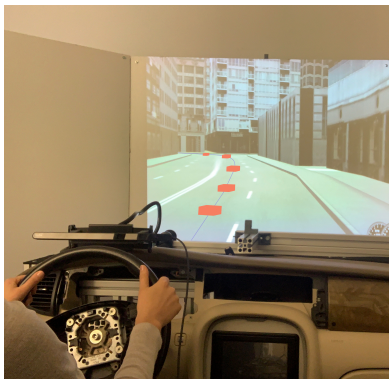
Figure 4.5: Schematic representation of the successive steps deployed in VR-scene development.

length and complexity were created (see Figure 4.4). On average, it took three minutes and thirty seconds to finish each route using the autopilot. Apart from one route segment that was shared by two maps, the maps did not overlap with each other. The shared segment was passed in opposite directions in the different routes so that carryover effects should be minimal. The environment of each map consisted of buildings, traffic lights, road signs, road markers, and pavements with bus stops, trees, and small parkways, resulting in a fairly rich and complex environment. The driving environment did not contain any pedestrians or traffic. In all three maps, the lighting conditions corresponded to diffuse skylight at noon. The lighting gave some information about the orientation of buildings with respect to the sun, but no strong shadows were cast. Figure 4.5 summarizes the procedural steps in the deployment of the virtual environment and route programming for the autopilot driving condition.

A Din A4 color print for each virtual map with a designated driving route was created, including start and finish points (see Figure 4.4). In the learning phase, the respective map was placed on the cockpit's dashboard where the participant could easily see

(**a**) Driving with autopilot

(**b**) Driving with map



(**c**) Driving with navigation system (red squares projected on the road)

Figure 4.6: Overview of the driving modes used in the learning phase.

and handle it (see Figure 4.6(**b**)). The navigation system in the learning phase was represented as a blue, continuous line with red squares projected on the road and going from the start to the end of the entire driving route (see Figure 4.6(**c**)). This projection was programmed separately for each route. Finally, the autopilot condition of the learning phase was designed as a passive drive on a coordinate-based pre-programmed route for all three virtual maps.

Landmarks

For each map, six unique virtual landmarks were selected. Several landmarks were downloaded from Sketchfab [189] and adopted to the virtual environment using jMon-

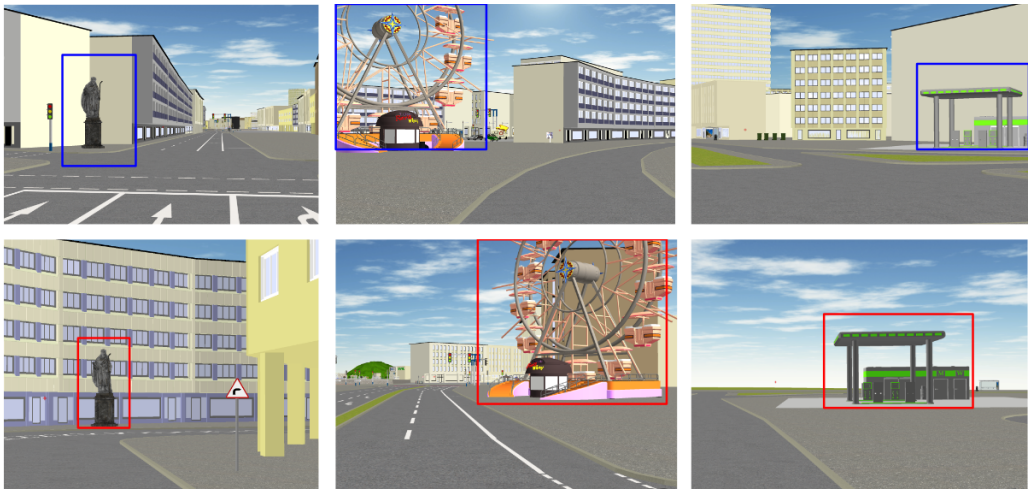


Figure 4.7: Examples of the landmarks and foils. **Top** (objects in the blue frame): the landmarks presented in the virtual environment. **Bottom** (objects in the red frame): foils used in the landmark recognition test.

keyEngine and Blender [25]. The landmarks were selected to be eye-catching so that they would be noticed and recalled by participants independently of the context [144] (see Figure 4.7). The landmarks were placed either along the driving route on the left or right side or next to an intersection of the roads. In the participant's instruction and test application (discussed below), the word "landmark" was deliberately avoided using "object" instead to exclude biasing the participant with respect to landmark-based vs. route-based learning strategies. In the experimental instruction, the word "scene" was used to describe the virtual environment and to avoid any confusion with a printed map.

Procedure

The experiment was approved by the local ethics board (TU Kaiserslautern). To avoid any order or learning effects, the order of the driving assistants and maps was randomized within subjects using the Balanced Latin Square method [123].

The experiment took place in a quiet room under controlled lighting conditions. After arriving at the laboratory, participants received a written informed consent form containing general information about the purpose, possible risks, and benefits of the study and also a sheet to permit the camera recording. After these forms were signed, general information was collected, including age, gender, visual acuity, and the driving

frequency per week. Subsequently, the participants received instructions about the experimental process. It was explained that they would drive three routes twice: once using a driving assistant, such as a printed map, autopilot, or a navigation system (learning phase), and once without using the assistant (test phase). Importantly, participants were aware that they would later have to drive the same route themselves by memory. After ensuring the participants had no questions regarding the experimental procedure, the familiarization phase was started, where participants learned to steer the driving simulator system in a training environment. The familiarization period lasted between ten and twelve minutes. The participants were instructed to follow general driving rules, including holding the proper traffic line, avoiding crashes with objects, and controlling their speed.

After the familiarization phase, the learning phase began. For the autopilot condition, the participants were instructed not to use the steering wheels and pedals, keeping the hands possibly on the knees. In the navigation system condition, participants were asked to follow the navigation projection as precisely as possible. In the map condition, participants were given unlimited time to learn the route from the map prior to driving. In the learning phase, they were also allowed to use the map for navigation. If participants had problems reading the map, the experimenter gave short hints to facilitate understanding. Deviations from the route while driving with the map were not counted as an error in the learning phase. As soon as the participants arrived at the destination, the map was removed from the cockpit. In the autopilot and navigation system conditions, the end of the route was indicated by the pop-up message "Destination".

Participants had to repeat the just-driven route in the test phase without the navigation aid. At this point, each deviation from the original route (e.g., wrong turn or skipping the right turn) was counted as a "wrong turn". If participants did not notice they were driving the wrong route within ten seconds, the experimenter shortly warned participants against the deviation and helped them to drive back to the position where the deviation from the right route had occurred. From that point, participants had to find the correct direction on their own.

Subsequently, the participants were informed about the NASA-TLX test conducted after every driving segment. The test was followed by the scene recognition and route knowledge tests after every first driving phase, and by the map drawing task after every

second driving phase. For every driving segment, the scene recognition test included six real scenes with correctly placed objects in them and six fake scenes (foils) containing the objects in the wrong places, resulting in twelve scenes in total (see Figure 4.7). The participants were informed about the number of foils and asked to select exactly six scenes. In the next step, route knowledge was tested by asking participants to put the selected scenes into the order they were encountered while driving. No feedback was provided either for the selected scenes' correctness or their order by the application.

The weighted NASA-TLX test, landmark recognition, and route knowledge task were programmed in JavaScript using the VUE framework and run on macOS Big Sur.

4.5 Statistical Analysis

To investigate the possible effects of the assistant type and the driving phase, a two-way within-subjects analysis of variance (ANOVA) was conducted. All reported p-values for the main and contrast effects are Greenhouse–Geisser corrected to safeguard against underestimation of p-values when the statistical assumption of compound symmetry is violated. To analyze simple effects within each factor, planned pairwise t-tests were employed with Bonferroni correction. All significant results are reported. A $p < 0.05$ threshold was used for the statistical significance. The data aggregation and signal processing were performed using Python; statistical analysis was run using R.

Cognitive Workload

Figure 4.8 represents the descriptive statistics for reported cognitive workload among the assistant types and driving phases. The ANOVA showed neither a significant main effect of assistant type [$F(2, 38) = 1.38$, $p = 0.61$], nor of the driving phase [$F(1, 19) = 1.38$, $p = 0.25$]. However, a highly significant interaction effect between these two variables was present [$F(2, 38) = 19.33$, $p < 0.01$]. Whereas for the autopilot and navigation system, cognitive load was lower in the learning phase than in the test phase $t(119) = -6.05$, $p < 0.01$ and $t(119) = -7.10$, $p < 0.01$, the reverse effect was observed for the printed map $t(119) = 9.69$, $p < 0.01$. This data pattern suggests that first investing cognitive effort when learning the environment by map lowers cognitive costs at retrieval. In contrast, the reduced cognitive effort when driving passively or by a

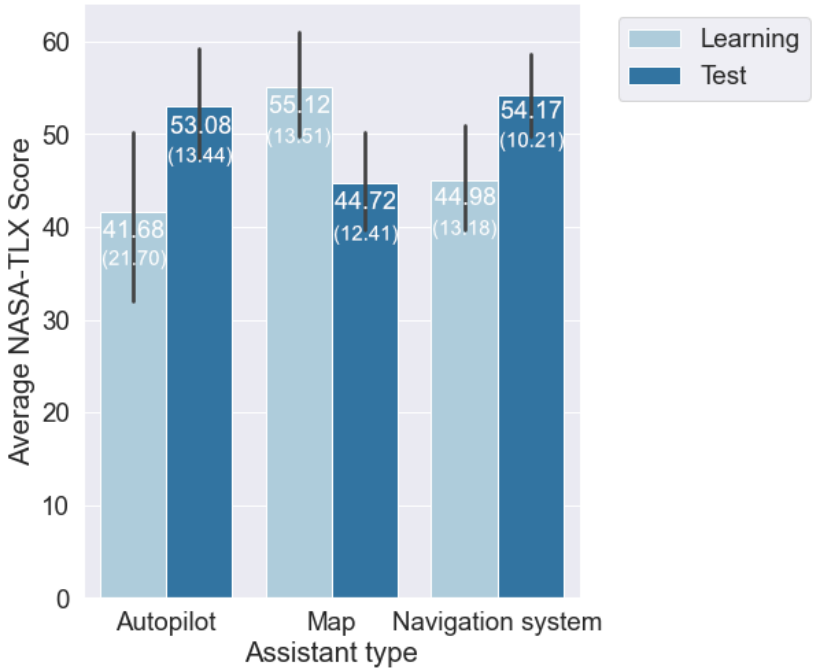


Figure 4.8: Descriptive statistics for NASA-TLX score by assistant type and driving phase: mean, standard deviation. Black vertical lines represents standard error. Learning a new route with a map was associated with the highest NASA-TLX score. In contrast, the lowest NASA-TLX score was observed in the test phase after learning a new route with a map.

navigation system comes at the expense of greater cognitive costs when those assistant systems become unavailable.

EDA

To extract underlying SCL and SCR components from the EDA signal, the Convex Optimization Approach (cvxEDA) was applied as proposed by Greco and colleagues [93]. This method is a reasonable choice when SCRs occur to multiple stimuli and overlap over the entire period of interest [46].

The raw EDA signal was filtered with an 8th order Chebyshev Type I low-pass filter (0.1 Hz) for all participants. In the next step, a Z-score normalization was applied to the filtered data. Then, cvxEDA was run on filtered and normalized data for all driving segments (see Figure 2.4). To avoid any residual motion artefacts in the signal,

Table 4.2: Descriptive statistics for the tonic and phasic components by the assistant type and driving phase. All values are Z-score normalized.

		Tonic		Phasic	
Assistant Type	**Phase**	**Mean**	**SD**	**Mean**	**SD**
Autopilot	Learning	−0.12	0.84	0.03	0.03
Autopilot	Test	−0.05	0.79	0.08	0.07
Navigation system	Learning	−0.20	0.85	0.06	0.06
Navigation system	Test	0.17	0.97	0.07	0.08
Map	Learning	−0.26	0.94	0.08	0.08
Map	Test	−0.35	0.75	0.05	0.04

the first 30 and the last 10 seconds from the signal of each driving segment were deliberately excluded. After tonic and phasic components were extracted from the EDA data, an average value was calculated for each participant within each driving segment. The data of one participant were excluded from the analysis (and the correlation analysis discussed below) because of insufficient recording quality.

In the next step, possible influences of the driving phase, assistant type, and their interaction on the EDA components were investigated. Table 4.2 represents descriptive statistics for the tonic and phasic components. For the tonic component, neither driving phase [$F(1, 18) = 0.95$, $p = 0.34$)], nor assistant type, [$F(2, 36) = 0.61$, $p = 0.54$], nor their interaction [$F(2, 36) = 1.54$, $p = 0.23$] were significant. In contrast, a statistically significant interaction effect between the driving phase and assistant type was observed for the phasic component [$F(2, 36) = 5.38$, $p < 0.05$]. There were no significant main effects of either driving phase or assistant type on the phasic component [$F(2, 36) = 0.17$, $p = 0.80$] and [$F(1, 18) = 2.82$, $p = 0.11$] respectively.

The pattern of results was similar to that of the NASA-TLX scores. A pairwise t-test yielded a significant difference between learning and test phases with the autopilot in the phasic component values, $t(18) = -3.19$, $p = 0.005$. The mean value of the phasic component was significantly lower in the learning phase with the autopilot, compared to the corresponding test phase (see Table 4.2). In the map condition, this effect was reversed and of similar size, but failed to meet the significance criterion, $t(18) = 2.06$, $p = 0.054$. No difference between the learning and test phases was detectable for the navigation system, $t(18) = -1.01$, $p = 0.33$.

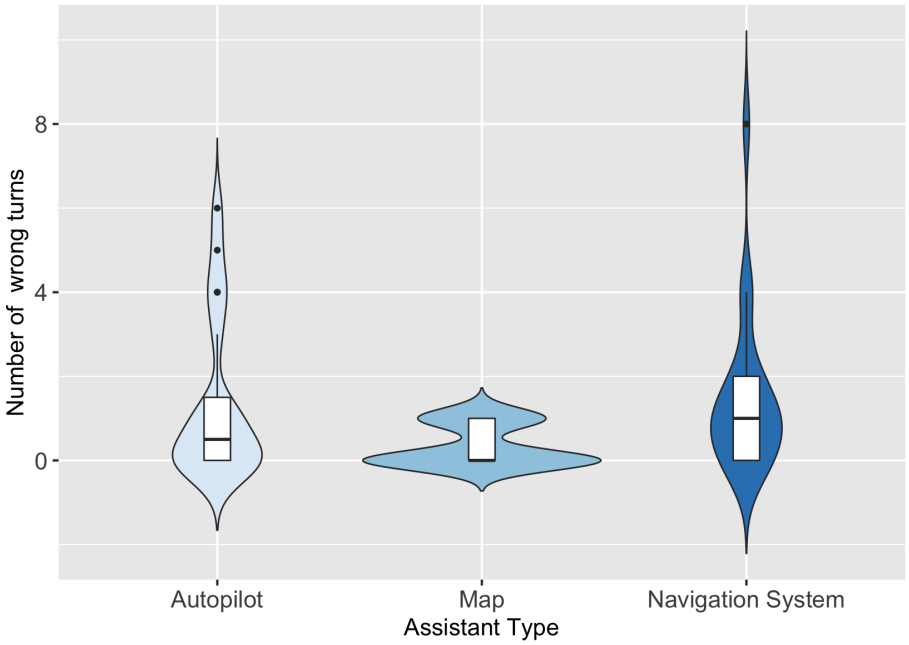


Figure 4.9: Distribution of the number of wrong turns for the different assistant types. Black dots represent outliers. Learning a new route with a printed map resulted in the lowest number of wrong turns in the test phase.

Landmark Recognition, Route Knowledge, and Way-Finding

To investigate possible differences in scene and landmark recognition among the three types of driving assistants, the proportions of ”hits” and ”false alarms” were calculated for each subject and transformed by arcsin transformation to meet ANOVA requirements. Route knowledge was calculated as a proportion of landmarks placed in the correct order and also arcsin-transformed. One-way ANOVA with the single factor assistant type showed neither any significant influence on hit rate [$F(2, 38) = 0.75$, $p = 0.48$], nor on false alarm rate [$F(2, 38) = 1.06$, $p = 0.36$], nor on the number of landmarks placed in the correct order [$F(2, 38) = 0.31$, $p = 0.74$].

Finally, the way-finding performance was evaluated. Way-finding performance was assessed through the number of wrong turns within the test phase (Figure 4.9). On average, each participant made a wrong turn 1.35 ($SD = 1.93$) times after learning with the autopilot, 0.35 ($SD = 0.49$) times after learning with the map, and 1.50 ($SD = 1.93$) times after learning with the navigation system. Due to pronounced outliers in

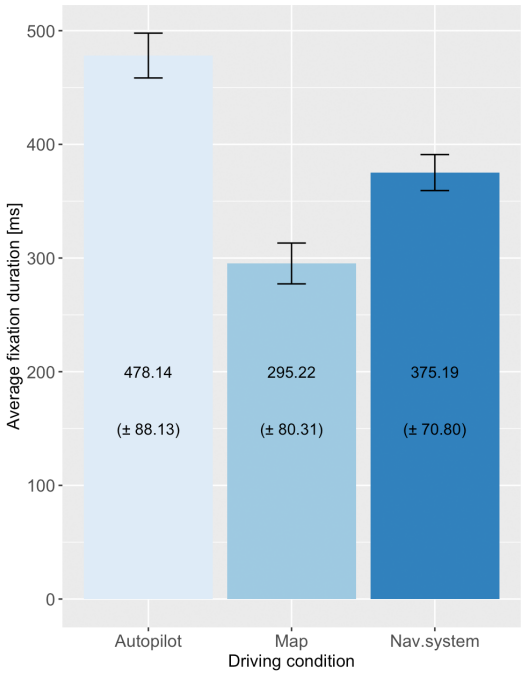


Figure 4.10: Histogram plot with mean values standard errors for average fixation duration by driving conditions. Standard deviations are represented in parentheses. Driving with autopilot resulted in the highest average fixation duration.

the number of wrong turns after learning with the autopilot or the navigation system, the normality assumption was violated (Shapiro–Wilk test $W = 0.67$, $p < 0.01$). Hence, the Friedman rank-sum test was employed to compare the group means. The results showed statistically significant changes in the frequency of making a wrong turn depending on the driving assistant in the learning phase, $\chi^2(2) = 6.23$, $p = 0.05$. Next, post-hoc Wilcoxon signed-rank tests with Bonferroni corrections were performed. The analysis confirmed a significantly lower number of wrong turns after learning with a printed map than after learning with the navigation system ($Z = 12$, $p = 0.02$). The difference between the autopilot and printed map conditions was not significant ($Z = 66$, $p = 0.1$). Similarly, there was no significant difference between autopilot and navigation system ($Z = 81.5$, $p = 1.0$).

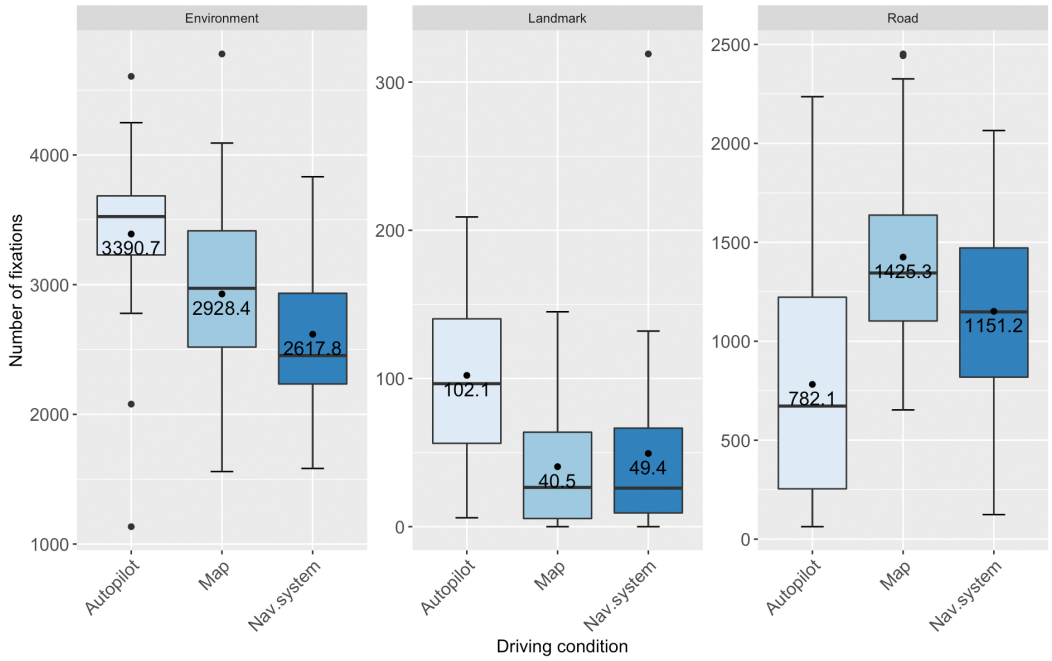


Figure 4.11: Box plots of the number of fixations grouped by the driving conditions and types of AOIs. The number in the box plots represents the average value within each driving condition and type of AOI. The black dots outside the box plots represent outliers. While driving with autopilot, participants primarly gazed at environmental objects and landmarks. In contrast, while driving with a map the most of fixations were associated with gazing at the road.

Eye Movement

To examine possible impact of navigation mode on eye-movement, eye-features were additionally analysed. Since the used three scenes were slightly different in geometries and in the number of curves and road intersections (but with similar overall complexity), only time-related and AOI-related features were analysed: *fixation duration, the total number of environment-directed gazes, the total number of landmark-directed gazes*, and *total number of on-road gazes*. The analysis was performed on the data averaged on the individual level and grouped by the driving condition. Data normality was examined using the Shapiro-Wilk test. The reported F-values are Greenhouse-Geisser corrected. Post hoc analysis was run using Tukey's HSD.

There was a highly significant main effect of driving condition on the fixation duration

[F(2, 38) = 26.24, p $<$ 0.001], with the longest fixations in the autopilot condition, followed by the navigation system and finally, the map condition (see Fig. 4.10). Post hoc Tukey's HSD showed that the average fixation duration was higher in the autopilot condition compared to driving with the navigation system [p $<$ 0.001] and also to driving with a map [p = 0.01]. No significant difference was observed between the map and navigation system [p = 0.60].

Next, AOI-related features (see Fig. 4.11) were analyzed. There was a significant main effect of driving condition on the total number of landmark-related objects [F(2, 38) = 4.96, p = 0.02], with a higher number in the autopilot than in the map condition [p = 0.05] and the navigation system condition [p = 0.02]. No differences were observed between driving with a map or navigation system [p = 0.87]. The number of on-road gazes also was significantly impacted by the driving condition [F(2, 38) = 12.88, p $<$ 0.001]. In particular, the number of on-road gazes in the map condition was significantly higher than in autonomous driving mode [p = 0.002]. Again, no significant difference was observed between the map and navigation system conditions [p = 0.29]. Finally, a significant main effect was observed for the total number of environment-directed gazes [F(2, 38) = 7.34, p = 0.003]. The post-hoc test showed that this difference was only significant between autopilot and navigation system conditions [p = 0.01].

Correlation Analysis

To explore possible linear associations between cognitive workload, the number of wrong turns and the EDA components, a multivariate product-moment correlation analysis Pearson's r was computed. Since the number of wrong turns was only recorded in the test phase, the NASA-TLX score and EDA components (tonic and phasic) from the learning phase were discarded from the correlation analysis. The correlation strength is reported following the classification by Dancey and Reaidy [12, 59].

A significant, weak, and positive correlation was observed between the number of wrong turns and the NASA-TLX score ($r = 0.32$, $p < 0.05$) (see Figure 4.12). Moreover, a highly significant, moderate, and positive correlation was obtained between the number of wrong turns and the phasic component ($r = 0.43$, $p < 0.01$). In the next phase, Pearson's r correlation analysis was run separately for the autopilot, navigation system, and printed map driving assistants using the same variables as before (see Figure 4.12).

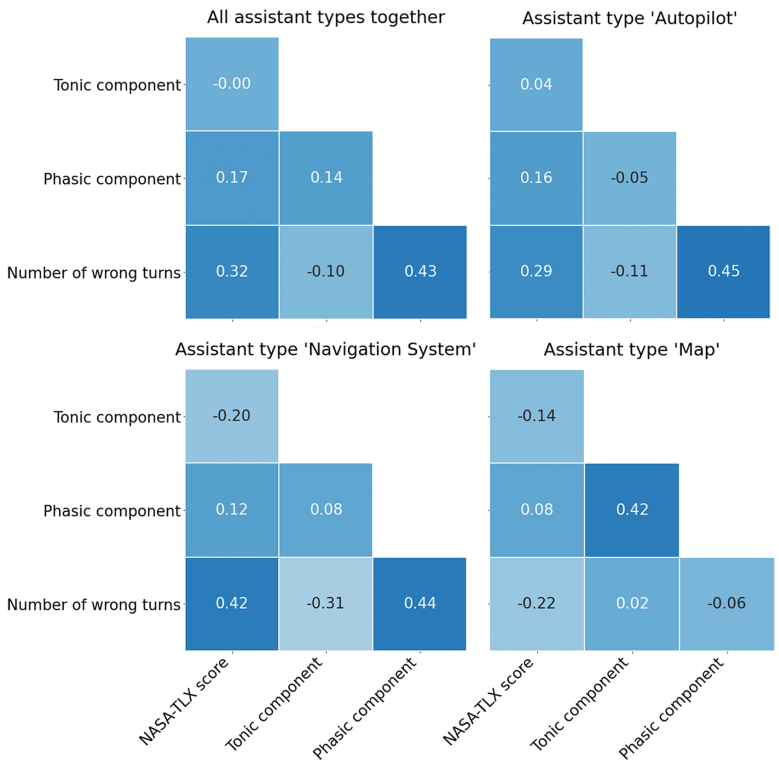


Figure 4.12: Correlation analysis. **Top left**: Test phase including all assistant types. **Top right**: Test phase for assistant type 'Autopilot'. **Bottom left**: Test phase for assistant type 'Navigation System'. **Bottom right**: Test phase for assistant type 'Map'.

For the autopilot, a statistically significant, moderate, and positive correlation between number of wrong turns and the average value of the phasic component ($r = 0.45$, $p = 0.05$) was found. For the navigation system, a similar correlation pattern was observed between the number of wrong turns and the phasic component, resulting in $r = 0.44$, $p < 0.06$, yet slightly exceeding the 95% significance level. For the map assistant type, no significant correlations were observed.

4.6 Machine Learning Approach

Machine learning algorithms provide a powerful tool for the detection and classification of eye movements both from raw and processed data [238]. In a way-finding study with 52 participants, Alinaghi and colleagues demonstrated the feasibility of the Gradient

Table 4.3: Overview of extracted and calculated statistical features for the eye movement.

Feature	Description	Features
Gaze X	Gaze FX X coordinate in Tobii's CS	μ, σ, RMS
FX SC X	Gaze FX X coordinate in screen CS	μ, σ, RMS
FX SC Y	Gaze FX Y coordinate in screen CS	μ, σ, RMS
FX contact point Y	Lateral Y position of FP in 3D space in the SCS	μ, σ, RMS
FX duration	FX duration in milliseconds	μ, σ, min, max, RMS
TT between objects	TT between 2 different objects in msec	μ, σ, min, max, RMS
FX 2D distance	Distance between 2 objects in 2D space	μ, σ, min, max, RMS
FX 3D distance	Distance between 2 objects in 3D space	μ, σ, min, max, RMS

CS: coordinate system, *FX*: fixation, *FP*: fixation point, *TT*: transition time, *SCS*: simulator coordinate system, *RMS*: root mean square.

Boosting algorithm for predicting drivers' turning decisions [13]. Using saccade and frequency-based features, they achieved an average classification accuracy of 86% across three types of behavior in a three-second window: turning left, turning right, and non-turning. Zahabi et al. ran a study to classify the driving situation of police officers (normal vs. pursuit driving) in a virtual environment [236]. Using the gaze data of 18 police officers (percentage change in pupil size and blink rate) along with driver behavior, the authors were able to achieve a classification accuracy of about 90% using Random Forest and Support Vector Machine (SVM) models. In a study on driver takeover performance, Du and colleagues evaluated the performance of six machine learning methods for the prediction of takeover performance (*good* or *bad*) [72]. Among other physiological features, the authors used fixations, saccades, pupil size, blink rate, gaze dispersion, scan pattern, and proportion of gazes focused on the road to build SVM, Random Forest, Naive Bayes, k-Nearest Neighbour (kNN), discriminant analysis, and logistic regression models. Random Forest showed the highest classification performance with an average accuracy of 0.82 and an F1-score of 0.62 using gaze and environmental features calculated within a three-second window. The second-best performance was demonstrated by SVM. Taking these studies together, it can be concluded that Random Forest, SVM, and Gradient Boosting algorithms provide robust techniques for classifying the driver's gaze in various scenarios. Alinaghi and colleagues demonstrated the feasibility of a Gradient Boosting algorithm for predicting

drivers' turning decisions [13]. Using saccade and frequency-based features, they achieved an average classification accuracy of 86% across three types of behavior in a three-second window: turning left, turning right, and non-turning. Zahabi et al. achieved an accuracy of about 53% using a Random Forest algorithm and gaze features to classify the driving mode (normal vs pursuit driving) in a virtual environment [236]. Du and colleagues evaluated the performance of six machine learning models to predict takeover performance. Random Forest-based model showed the highest classification performance with an average accuracy of 0.82 and an F1-score of 0.62 using gaze and environmental features calculated within a three-second window.

The limitations of the aforementioned studies and currently available eye-tracking systems in serial vehicles [149, 128] are twofold. First, the driver's engagement was primarily considered as a function of the current gaze position (off- or on-road). Nevertheless, this binary splitting of the gaze location cannot guarantee real engagement and proper perception of the situation on the road, including avoidance of some crash and near-crash risks [209, 134, 221, 102]. Second, the driving situations employed mainly were restricted to straight highway driving [33].

4.6.1 Feature Engineering

Table 4.3 overviews fixation-related features extracted from the OpenDS Gaze Analyzer. Statistical features for each gaze feature were calculated using sliding windows without overlap with frame lengths of four and ten seconds [13]. The calculation was carried out on the level of single participants for each driving task. OpenDS Gaze Analyzer additionally provides information related to the objects that were gazed at, e.g. buildings, streets, and landmarks. All the objects were split into two groups of area of interest (AOI): 1) environment and 2) on-road. For both groups, the percentage of the occurrence within a sliding window was estimated. In total, 33 features were used to train the machine learning models.

4.6.2 Model Building

Both binary and multi-class models were considered. For the multi-class models, all three driving conditions (*autopilot*, *map*, and *navigation system*) were considered as

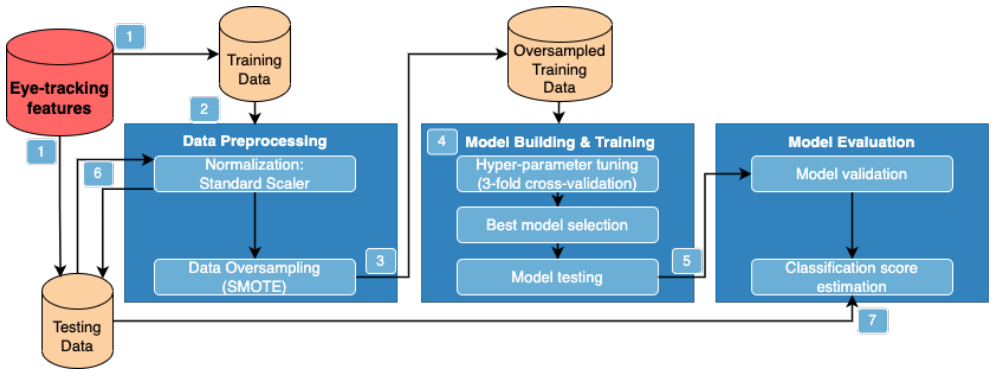


Figure 4.13: Model building and evaluation pipeline for driving mode classification using eye-tracking features.

separate classes. In the case of binary classification, two driving conditions (*map* and *navigation system*) were combined into one class (*manual driving*) and contrasted with the class *autopilot*. It is hypothesized that the driver's engagement level with driving is high in manual driving conditions and low in the autopilot condition.

Before being fed into the classification models, the data was processed in two ways: At the beginning, StandardScaler was applied from the Scikit-learn library to achieve equal scaling among input features. Next, the class distribution in the data set was not equal. For the binary models, the distribution of class labels resulted in 38.65% for class 1 (*autopilot*) and 61.35% for class 2 (*manual driving*), respectively. For the multi-class models, the label distribution was balanced, resulting in 38.65% for class 1 (*autopilot*), 30.17% for class 2 (*map*) and 31.18% for class 3 (*navigation system*). However, after creating training and testing data sets, the minority classes in binary and multi-class models were over-sampled using the synthetic minority over-sampling technique (SMOTE). Figure 4.13 summarizes the model building and evaluation pipeline. The SMOTE technique was applied only to the training data sets.

A nested cross-validation procedure was used to achieve a robust classification performance and avoid the risk of model overfitting [48]. In the inner loop, the GridSearchCV method was deployed to define a grid of hyper-parameters and perform 3-fold cross-validation. The model evaluation was performed 20 times on the validation data following the total number of participants.

To cope with the computational costs increasing with the number of hyperparameters, the definition of the parameter set and search space was partly adopted from [238] for

Table 4.4: Hyper-parameter search grid with the lower and upper boundaries for the used values.

Algorithm	Hyper-Parameter*	Lower	Upper
Random Forest	d_{max}	10	60
	f_{max}	1	18
	l_{min}	1	4
	s_{min}	2	10
	n	2	180
	Bootstrap	{False}	
	Criterion	{Gini}	
Gradient Boosting	n	200	800
	λ	0.1	0.5
	s	0.1	0.5
	w_{min}	0.0	0.1
	d_{max}	2	16
	Early stopping	{10}	

*d: depth, f: features, l: sample leaf s: sample split

*n: number of estimators, λ: learning rate, w: weight leaf

Table 4.5: Classification results of the Gradient Boosting and Random Forest algorithms using features calculated within a 4-seconds window. The standard deviation for each metric is denoted in parentheses.

Model	Classes	4 seconds		
		Accuracy	**AUC**	**F1-score**
Random Forest	2	0.893 (0.047)	0.945 (0.047)	0.891 (0.051)
	3	0.744 (0.074)	0.893 (0.056)	0.741 (0.076)
Gradient Boosting	2	**0.901 (0.043)**	**0.952 (0.051)**	**0.900 (0.046)**
	3	0.742 (0.075)	0.900 (0.060)	0.738 (0.077)

Random Forest and from [13] for Gradient Boosting. Table 4.4 lists the tuned hyper-parameters and their value ranges. The results are reported using F1-score (weighted average) [95], average accuracy [198] and the area under the Receiver Operating Characteristic curve (AUC ROC).

4.6.3 Results

Tables 4.5 and 4.6 represent the classification results obtained for features calculated within 4 and 10-second windows. The highest classification performance was achieved

Table 4.6: Classification results of the Gradient Boosting and Random Forest algorithms using features calculated within a 10-seconds window. Best results are highlighted in boldface. The standard deviation for each metric is denoted in parentheses.

Model	**Classes**	**10 seconds**		
		Accuracy	**AUC**	**F1-score**
Random Forest	2	0.883 (0.050)	0.946 (0.057)	0.880 (0.053)
	3	0.775 (0.075)	0.916 (0.056)	0.770 (0.078)
Gradient Boosting	2	0.890 (0.074)	0.945 (0.0479)	0.889 (0.075)
	3	**0.803 (0.077)**	**0.927 (0.050)**	**0.800 (0.079)**

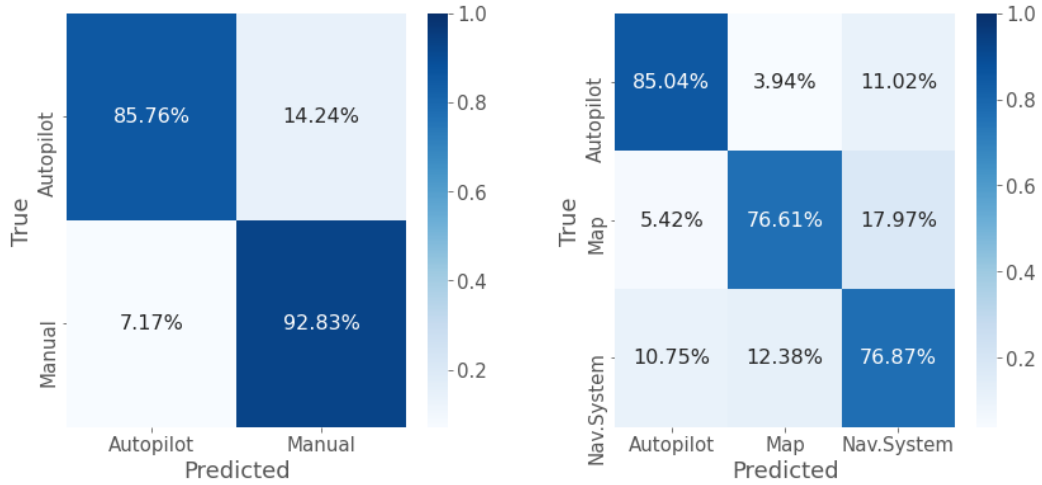


Figure 4.14: Confusion matrices for autonomous vs. manual binary and multiclass classification.

with the Gradient Boosting algorithm. As it can be seen, the highest classification performance for two classes was achieved with the features calculated within four-second windows. For three classes, the highest classification performance was achieved using a window size of ten seconds resulting in an accuracy of 0.8 ($\pm$0.07), F1-score of 0.81($\pm$0.08), and AUC of 0.93 ($\pm$0.06). Figure 4.14 additionally presents confusion matrices for both binary (four-second window) and multi-class (ten-second window) classification performance. In the multi-class classification, the largest confusion occurred between the classes *map* and *navigation system*. The lowest confusion was observed between the classes *autopilot* and *map*.

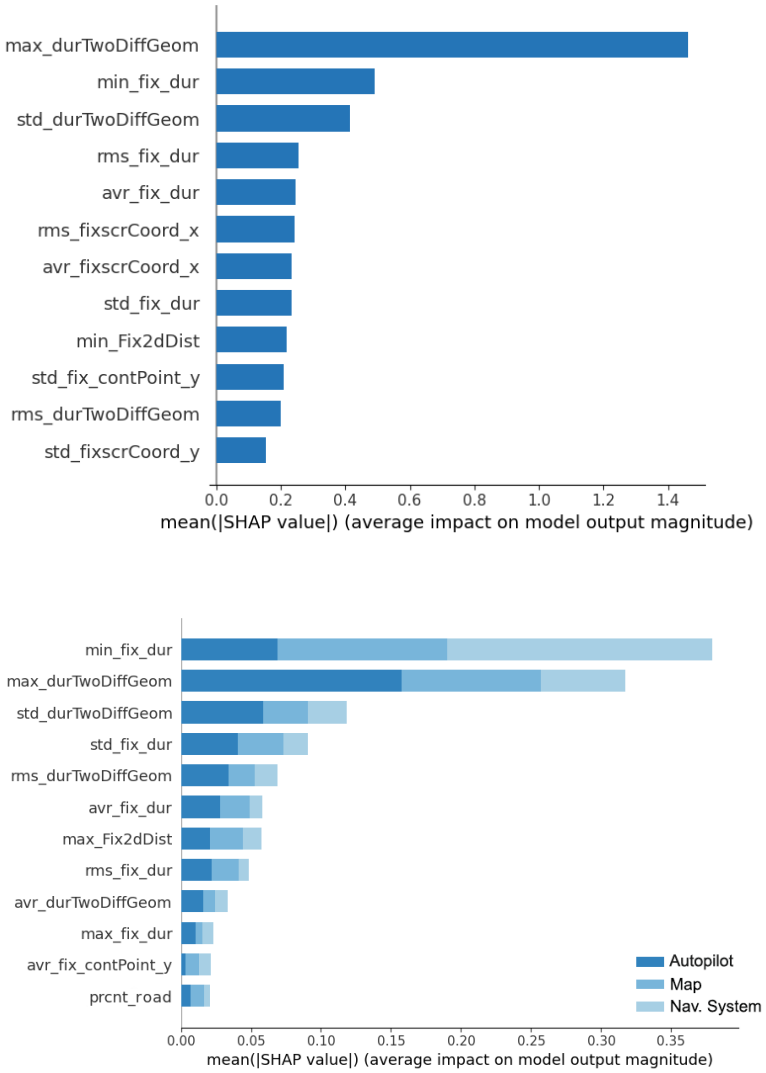


Figure 4.15: Feature importance graph for the Gradient Boosting classification models using the SHAP method (top 12 features). **Top**: SHAP values for binary model. **Bottom**: SHAP values for the multi-class model.

4.6.4 Model Explainability

To evaluate the contribution of individual features to the model classification performance, the SHAP (Shapley Additive Explanations) method was deployed. In this

method, each feature receives an importance value for a particular prediction [140]. Figure 4.15 represents SHAP values for binary (four-second window) and multi-class (ten-second window) classifications. Interestingly, the top three most important features for both binary and multi-class classification models are *max transition time between two objects*, *min fixation duration*, and *standard deviation in the transition time between two objects*. Remarkably, the AOI features relating to "off"- and "on"-road gazes did not appear highly important for the classification performance.

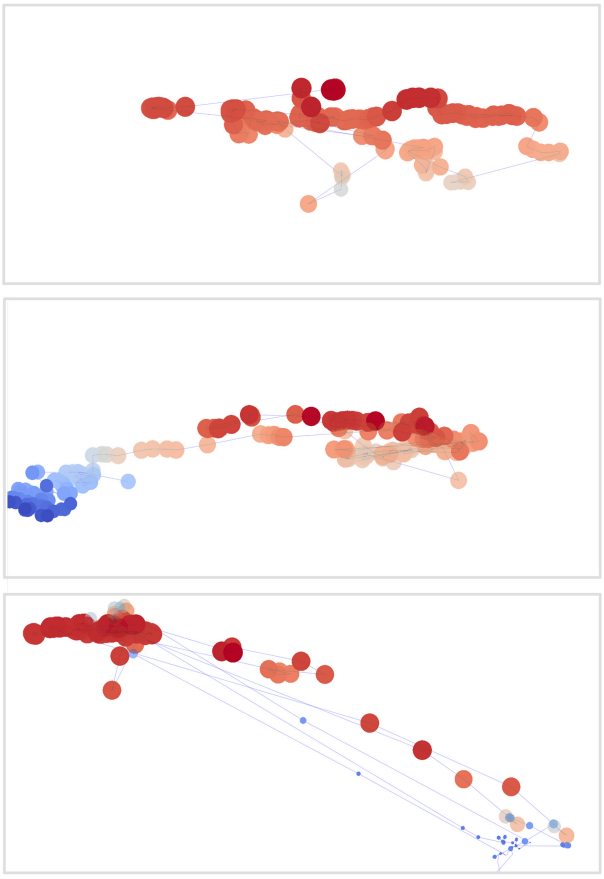


Figure 4.16: Randomly selected scan paths of ten-second length in three navigation conditions from the same participant. The red color represents long fixations, and the blue short ones. **Top**: Driving with the autopilot. **Middle**: Driving with the navigation system. **Bottom**: Driving with the map.

Figure 4.16 shows the scan path plots with respect to the three driving conditions. Driving with the autopilot was characterised by dense, long-lasting fixations distributed

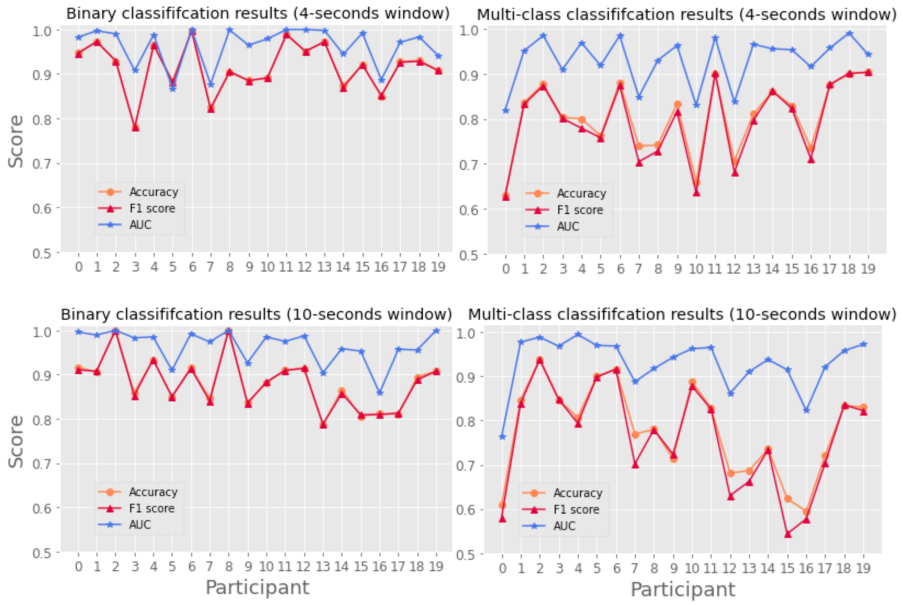


Figure 4.17: Classification performance of user-dependent binary and multi-class models by window size.

horizontally across the screen center, resembling a scanning behaviour. Driving with the navigation system showed a high variation in fixation duration which could be primarily attributed to fixating the navigation trace provided by the navigation system. Driving with the map showed two clusters of fixations: long-lasting, environment-related fixations on the top left and sparse, short ones in the bottom right corner where the printed map was placed.

User-dependent models

User-dependent models were built to examine whether a higher classification performance could be achieved by training and evaluating data from the same person. The set of parameters that performed best in the user-independent modality was used. In particular, for the binary classification, the best set of parameters was: $\lambda = 0.2$, $d_{max} = 4$, $f_{max} = 9$, $w_{min} = 0.0$, $n = 600$, $s = 0.4$, and for multi-class $\lambda = 0.2$, $d_{max} = 8$, $f_{max} = 7$, $w_{min} = 0.0$, $n = 600$, $s = 0.4$.

The models were built and trained (70% training split and 30% test split) on the level of the individual 10 times to address the uncertainty within training and test splitting

(i.e. 10 different random seeds). Accuracy, F1-score and AUC were computed for each test split. Thus, the final individual scores are the average value from the ten respective iterations.

Figure 4.17 represents user-dependent classification results by class type and window size. Both for binary and multi-class classification models, the highest accuracy was achieved using 4-second windows. On average, the individual models achieved an accuracy of 0.92 (± 0.06), F1-score of 0.92 (± 0.06) and AUC of 0.96 (± 0.04) for binary classification, and an accuracy of 0.78 (± 0.10), F1-score of 0.76 (± 0.12) and AUC of 0.93 (± 0.06) for multi-class, respectively. The drastic reduction in the amount of data available for training could explain the comparably low user-dependent classification performance for the multi-class case.

4.7 Conclusion

The presented study sought to understand spatial imagery development and provide a prospect towards an engagement-ware system under manual and autonomous driving conditions. For this purpose, spatial knowledge, mental workload and gaze data were accessed. The results demonstrate a considerable difference in spatial knowledge, cognitive workload and gaze-related features. Thus, initial learning of a new environment through a printed map results in a higher cognitive demand that significantly diminishes when driving the same route by memory. In contrast, this effect was the opposite for the navigation system and the autopilot, for which the initially low cognitive workload in the learning phase dramatically increased in the test phase. Moreover, after learning the printed map, participants tended to drive the route in the test phase with fewer errors. It is also shown for the first time that the phasic component of the skin conductance response parallels behavioral measures so that EDA measurement may serve as a possible proxy for self-reported cognitive workload or stress. Despite minor expectations, driving with the autopilot was not associated with benefits in landmark recognition.

The observed results demonstrate that driving with autopilot, a navigation system or a printed map produces distinct eye movement patterns that can be classified with considerable accuracy using machine learning algorithms. Random Forest achieved comparable results for both binary and multi-class classification performance. In line

with previous works, this performance was influenced by the window size selected for the feature calculation [13]. For the binary models, the four-second window size showed the highest classification performance, whereas, for the three classes, the performance was higher with a ten-second window. The explanation could be two-fold. First, considering the feature importance obtained by the SHAP value method, it is assumed that the minimum and maximum values could carry more information when calculated within broader ranges (i.e. within larger windows), thus allowing for better discrimination among the three classes. Second, when calculating features within four-second windows, the noise ratio might remain high, which could drastically restrict the discrimination power for multiple classes. The Gradient Boosting algorithm achieved the highest classification accuracy of 90.1 % for binary driving mode prediction, (*autopilot* vs *manual driving*), and an accuracy of 80.3% for multi-class classification, (*autopilot*, *map* and *navigation system*). The classification performance varied between window sizes similarly for both algorithms. Comparing this classification accuracy to the state of the art is challenging since the related studies vary in experimental design and used methods. Yet, as in the study by Alinaghi et al. [13], the Gradient Boosting achieved the highest classification performance compared to Random Forest. Moreover, our binary prediction accuracy of the driving mode exceeded those of Zahabi et al. [236] by 37, and those of Du et al. by 8 percentage points.

Additionally, the contribution of single features to the classification performance were evaluated. The *min fixation duration* showed the highest importance in the multi-class classification performance, especially for the classes *navigation system* and *map*. The short fixation duration could be explained due to the fact that participants used the on-screen navigation trace, for which rapid fixations were sufficient to estimate the driving direction. In the case of the map condition, the drivers had to permanently compare their position with the printed map and simultaneously update their referential position in the driving environment. Indeed, the decrease in fixation duration could mean that the environmental objects receive less attention [49]. This assumption is partly supported by the post-experimental interviews performed in the original study, where some participants stated that they could not observe the environment because they had to concentrate on the prescribed route. The *max transition time between objects* was the second most important feature in the multi-class classification. While driving with the autopilot, drivers used a visual scanning strategy that requires a vision with

high acuity and excludes long saccades, which suppress sharp scene perception [164]. Considering these results, it can be concluded that eye-tracking technology provides sufficient information about the drivers' ongoing engagement and can serve as a solid basis for engagement-aware assistant systems. Based on the recognized engagement level, the system could adapt its warning behavior to keep the driver in the loop or even to take the vehicle out of the traffic in case the human driver is detached from driving and can pose a risk to other traffic participants.

This study has several limitations. First, the study design considered only a short-term impact of the driving assistants when a route was driven only twice. Thus, the long-term impact of the driving assistant types is still to be investigated. Second, a relatively small and simple environment was used without complex turns, road forks, traffic, and pedestrians. The use of a more complex and dynamic environment could bring more insights into spatial learning and cognitive workload of drivers under different navigation modes.

The final remark relates to the application of this study. The reduction in drivers' mental workload through the utilization of the navigation system and the autonomous driving mode was shown to occur at the cost of spatial learning. This should be considered in the future when designing novel navigation and transportation means as well as strategies to hand over vehicle control from the autonomous system to the driver, e.g., when a complex navigational decision has to be made, the technology fails, or a detour needs to be improvised that the navigation system is not prepared for. The goal of modern navigation aids should be, therefore, not only to enhance the driving and navigation experience but to avoid detrimental effects on the driver's acquisition of spatial knowledge.

5 Radar-based Engagement - Aware System

5.1 Proposed Study

Considering the safety issues, including driver's disengagement from the driving loop under long-term autopilot utilization on the one hand, on the other hand existing limitations of eye-tracking systems on vehicle board, this chapter proposes a novel driver monitoring approach using a UWB Radar. A Radar-based systems allow to track the driver's upper torso and, in contrast to the camera, provide unbeatable benefits in terms of privacy. This chapter introduces a study where six activities associated with conventional, autonomous, and distracted driving were recorded in the context of engagement-aware systems. The study was performed under simulated driving conditions because of safety issues for the driver and passengers and currently restricted legal utilization of the autopilots under local law.

Using a Convolutional Neural Network (CNN), Pyramid Vision Transformer (PVT-Tiny), and a Long Short-Term Memory neural network (LSTM), the generalization ability of the network is evaluated in comparison to the prevalent practice of random stratified data splitting versus the more strict leave-one-participant-out cross-validation method. Doppler data were normalized with a simple interquartile range (IQR) normalization method, avoiding extensive pre-processing steps which might heavily depend on the used radar system [55]. This normalization method ensures the system's real-time application capability and enhances the method's transferability among different radar systems. Next, two data augmentation techniques were applied and evaluated on the obtained dataset to facilitate the generalization ability of the deep learning models. Finally, in contrast to prior work on the topic, access to the dataset acquired in this study

is provided to encourage comparison and enable the reproducibility of results. While several radar datasets in HAR and the healthcare domain are available [11, 83, 27], to the best of the thesis author's knowledge, so far, there are no publicly available radar datasets consisting of driving activities. The results of this work were partially published in Brishtel et al. [36].

5.2 Related Work

In-Cabin Diver Monitoring

Under the increasing level of automation available in production vehicles, continuous driver monitoring becomes a crucial safety factor [114]. To ensure drivers remain undistracted in the driving loop and to prevent a negative impact of the autonomous system on the ability of drivers to take over [89, 231, 134, 38], multiple methods from various research fields, including human-machine interaction, psychology, computer science, and ergonomics have been investigated. Thus, in-cabin driver monitoring cameras [120, 143] and eye-tracking systems were tested [221, 139, 237, 44, 89] and partly integrated into production vehicles. Each of these technologies has its strengths and weaknesses. For instance, despite the high precision of distraction recognition [206, 222], in-cabin cameras are considered by many drivers as an intrusion into their privacy. Furthermore, eye-tracking systems cannot fully infer the drivers' engagement in the driving loop even if their eyes are directed on the road [221], despite their overall great ability to discriminate among different driving modes [37]. As mentioned in Section 5.1, radar technologies offer versatile advantages in resolving the issues of conventional monitoring solutions. Despite the rising interest in radar in the context of in-cabin driver monitoring, the current state-of-the-art comprises only very few publications capturing this application area.

The potential of pulse ultra-wideband radar for in-cabin driver health monitoring and smartphone utilization was demonstrated in a study by Leem et al. [130]. The authors described the pre-processing and reconstruction of the leaking breathing pattern under different driving activities. They also introduced an algorithm to detect drivers' smartphone usage, pointing at radar technology as a potential technique for preventing car crashes. Similarly, Ding et al. used an FMCW radar to detect inattentive

driver behavior [66]. The authors ran a series of experiments in a real car environment, where the drivers performed seven different activities, including head flexion, rotation, and shaking, as well as body movement, sleepy behavior, and picking up a smartphone. Using range-Doppler maps, they extracted a new activity representation called a dynamic-Doppler trajectory (DRDT) map. Then, the associated activities from the DRDT range of interest, Doppler energy change, and dispersion features were extracted and used to build machine learning algorithms. Using decision trees, SVM, KNN and ensemble classifiers, the highest average accuracy they achieved for the task of in-cabin activity classification was 95%. It is important to note that the recorded activities primarily considered head motions, flexion, and rotation.

Machine Learning for HAR using Radar

Several studies demonstrated the feasibility of CNNs' model to handle radar data. Thus, using a pre-trained and fine-tuned ResNet-18 and simulated micro-Doppler spectrograms, Du et al. [71] achieved an average accuracy of 97.92% for six classes, including walking, boxing, crawling, jumping, and standing. Shao et al. [183] recorded six participants performing similar actions as in the previous work using a UWB radar. Creating a simple CNN model and using only range information for model training, they reached an average accuracy of 95.24% for activity recognition. However, their validation dataset resulted from a random splitting of the data on the level of individual samples and not participants. Using a dataset with 1633 micro-Doppler spectrograms relating to six classes, including falling, Taylor et al. [205] evaluated six different machine learning models. They showed that CNN (combined with PCA) achieved the highest classification accuracy of 95.30%

Considering radar data as time series with time-varying properties, several authors proposed LSTM-based classification approaches for HAR. Using raw spectrograms of six obtained activities (walking, sitting down, standing up, picking up an object, drinking water, and falling), Taylor et al. [205] reported an average accuracy of 80.48% for Uni-LSTM and 83.53% for Bi-LSTM. Noori et al. [153] classified five activities (lying, sitting on the bed with the legs on the bed, sitting on the bed with the legs on the floor, standing, and walking) obtained from 13 participants using a UWB radar. Using an Enhanced Discriminant analysis with LSTM, they achieved an average

classification accuracy of 99.6%. However, after applying the leave-one-out cross-validation strategy, the overall classification performance dropped to 66%. Li et al. [133] investigated a bi-directional LSTM approach for HAR. They used six activities (walking, running, jumping, boxing, standing, creeping) from the MOCAP database [213] to build an LSTM model. Their bi-directional LSTM achieved 90.3% accuracy. They also evaluated the impact of the sequence length on the classification performance and found a length between 0.6 and 1 second to be sufficient for the optimal classification performance [133].

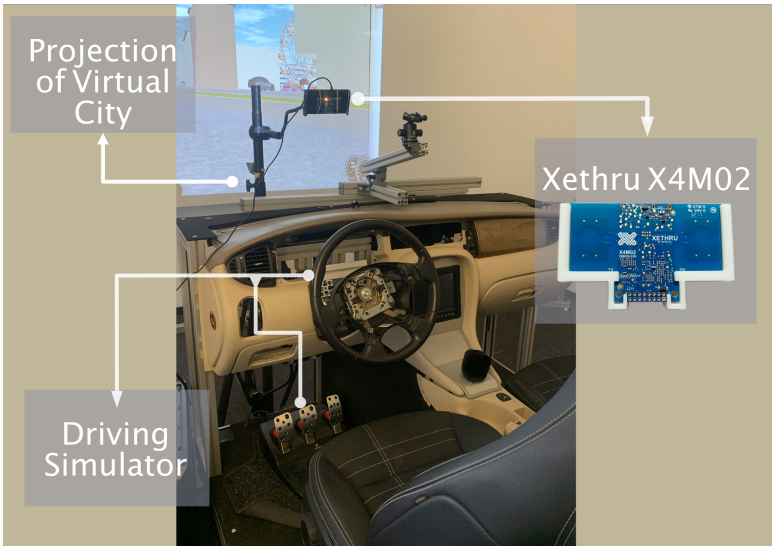


Figure 5.1: Setup of the recording environment using driving simulator with UWB Radar Xethru X4M02. Adopted from Brishtel et al. [36].

The unique nature of radar data to be represented as image and time-varying signals renders them highly attractive for Transformer networks. By now, however, only very few works on HAR with Radar have used a Transformer-based network for classification tasks. Qu et al. demonstrated the feasibility of a pre-trained Vision Transformer (ViT) for a hand gesture classification task recorded by an FMCW radar [169]. They recorded four classes, representing swipe, click, pinch and wave. Compared to 74.2% accuracy achieved by their SVM quadratic classifier, the ViT obtained an average accuracy of 97.5%. Wang et al. evaluated the performance of three ViT networks for human activity classification tasks and compared them with those of CNN and LSTM [227]. They

extracted Micro-Doppler Maps with five activities: walking, running, squat and stand, bowing, and turning around, recorded by a millimeter Wave (mmWave) radar. The highest classification accuracy was obtained by the introduced modification of the Vision Transformer-slice model with 99.12 %. In contrast, the average classification accuracy for a CNN and LSTM reached 95.23 % and 92.68 %, respectively. Using a publicly available dataset captured by an FMCW radar, Dey et al. introduced ViT-modifications for binary fall detection based on the magnitude of range time plots, range-Doppler plots and Micro-Doppler signatures [63]. Their modified ViT with Shifted Patch Tokenization and Locality-Self Attention achieved the highest classification accuracy of 97.54% on Micro-Doppler signature data. The model validation was performed using leave-one participant-out cross-validation method. In contrast, ResNet-50 demonstrated the highest accuracy of 89.02% in the same setting.

Most studies deploying radar systems for HAR reported outstanding classification performance of their machine learning models: in some cases, multi-class classifications exceeded an average accuracy of 90% [183, 153, 231, 205, 66, 227]. However, a detailed examination of these studies raises several questions regarding the generalization ability of the models. In particular, radar data acquired from multiple persons are commonly split randomly into training and test datasets [231, 153, 183, 205, 227]. Consequently, data from the same participant can be seen by the model during training and validation. Given the general ability of radar for biometric authentication [178, 131], this data splitting technique can not adequately evaluate the ability of the model to generalize to new users. Another area for improvement is the limited availability of radar datasets, which is crucial for reproducing reported results.

Considering previous studies' outcomes, both visual and time-varying representations of radar data perform on a very high level in human action recognition tasks. At the same time, only a few works [153, 219, 187] reported classification results from datasets with random stratified data splitting and cross-validation, where a significant difference was observed. Notably, the studies mentioned above used regular walking, crawling, standing up, sitting down, or falling, where each action has unique, clearly distinguishable patterns.

Table 5.1: Technical settings of Xethru X4M02 used for data recording.

Parameter	Value
Bandwidth (GHz)	7.25 - 10.20
Frames Per Second	50
Doppler Samples	1024
Doppler Frequency Range (Hz)	-8.5 - 8.5
Range Bins	24
Measurement Range (m)	0.4 - 1.20

5.3 Dataset Generation

5.3.1 Radar

The ultra-wideband (UWB) radar respiration sensor XeThru [5] X4M02 was used in the proposed study, which can detect and monitor human movements within the operating detection range [177]. Its low power consumption allows it to be integrated into portable devices. Table 5.1 lists the radar settings used for data recording. The radar placement was carried out following the empirical evaluations of Thullier et al. [210]. The detection zone was set to 0.40 to 1.20 m; the sensor was placed at a height of 60 cm over the cockpit, directed at the center of the driver seat (see Figure 5.1) to minimize obstacles and interference. It corresponded to a radar placement at the top of the windshield in a real car.

Because the default software for data recording does not provide any option for changing sampling frequency, the library *ModuleConnector* [154] was used to develop own script for recording and extraction of Doppler data. The radar data were sampled with an extended frequency of 50 fps. Due to the internal buffering process [5] of the Xethru radar, the resulting Doppler data had a frequency of 2.9 fps. Pulse-Doppler data was acquired containing the pulse magnitudes for all range bins and range values in the measured domain as well as the Doppler frequencies.

5.3.2 Driving Environment

The dataset was acquired in a mounted driving simulator consisting of a Jaguar XJ 4.2 V8 Executive cockpit and the integrated input controller Logitech G27 Driving Force comprised of a steering wheel, throttle, and brake pedals. The highly immersive driving simulation software OpenDS [156] was used to enhance realistic driving behavior for the participants. All driving tasks were performed using an automatic transmission.

Table 5.2: Overview of the data extracted from *RaDA*. Each sample contains a one-second window from a particular driving action.

Nr.	Action	Nb. of samples
1	Driving	1747
2	Autopilot	1844
3	Sleeping	1708
4	Driving & phone utilization	1692
5	Phone utilization	1715
6	Talking to passenger	1700
Total size		10.406

5.3.3 RaDA Dataset

Ten participants (one female) were asked to perform six activities as introduced in Table 5.2 and shown in Figure 5.2. Each participant performed the activities in the same fixed order. Each activity was recorded separately in a continuous manner. The total recording duration for each activity was set to one minute (minor deviations exceeding one minute are possible). Thus, the provided dataset includes approximately 60 minutes of driving activities. Table 5.3 provides information about the height and weight of participants included in the dataset.

5.3.4 Action Performance Protocol

Figure 5.3 represents the Doppler spectrograms for each class. Figure 5.2 provides graphical representations of the six recorded classes. Table 5.4 summarizes the six

Figure 5.2: Overview of six driving activities recorded with UWB radar Xethru X4M02. **Top** (left to right): Driving, Autopilot, Sleeping. **Bottom** (left to right): Driving & Smartphone Utilization, Smartphone Utilization, Talking to Passenger.

actions and their descriptions used during data collection. As mentioned earlier, the driving activities belong to three driving behaviours: normal or (engaged) driving (*driving*), autonomous driving (*autopilot*) and distracted driving (residual classes), where the classes *sleeping* and *smartphone utilization* are considered as distracted behaviour during autonomous driving. The definition of distracted driving behaviour was adopted from [173].

As motivated in Section 2.2.3, a convolutional neural network ResNet-18, an LSTM and a Transformer network PVT-Tiny were used to classify six different driving activities recorded with a UWB radar. The radar hardware automatically outputs the pulse-Doppler data after internally performing the fast Fourier transforms (FFT) on the time domain samples. Thus, the elements of the range-Doppler map are defined as $w_{k,f} \in \mathbb{R}$, where w denotes the Doppler pulse for a given range bin k, and Doppler frequency f. Range-Doppler maps W_t are generated at each frame measurement t.

Table 5.3: Weight and height of participants in the *RaDA* dataset.

Participant	Height (cm)	Weight (kg)
1	188	85
2	169	50
3	178	64
4	180	93
5	178	90
6	167	74
7	172	55
8	179	77
9	170	63
10	164	59

5.4 Machine Learning Approach

5.4.1 Data Extraction and Preprocessing

Based on empirical studies that showed two seconds of distracted behavior are sufficient for an increased risk of accidents [135, 221], a window size of 1 second was selected. Before being processed by deep learning models, the acquired data was cleaned from outliers using the interquartile range (IQR). As outliers were considered data points exceeding the range: *Q3 + 1.5 * IQR*, where Q3 is the third quartile (or 75^{th} percentile). Excessive amplitude values are caused by strong reflections from metallic objects, e.g., in parts of the car seats. The values exceeding this range were replaced by the maximum value within this range. Importantly, the IQR coefficient was calculated on the training data only and applied for data normalization in training as well as validation.

The following section describes the performed experiments for in-cabin driver activity classification. First, the baseline definition is introduced based on the re-implementation of the work of Ding et al. [66]. Then, the performance of ResNet-18 and an LSTM is investigated.

Table 5.4: Action performance protocol used for RaDA data acquisition. 10 participants performed all actions sequentially one minute long.

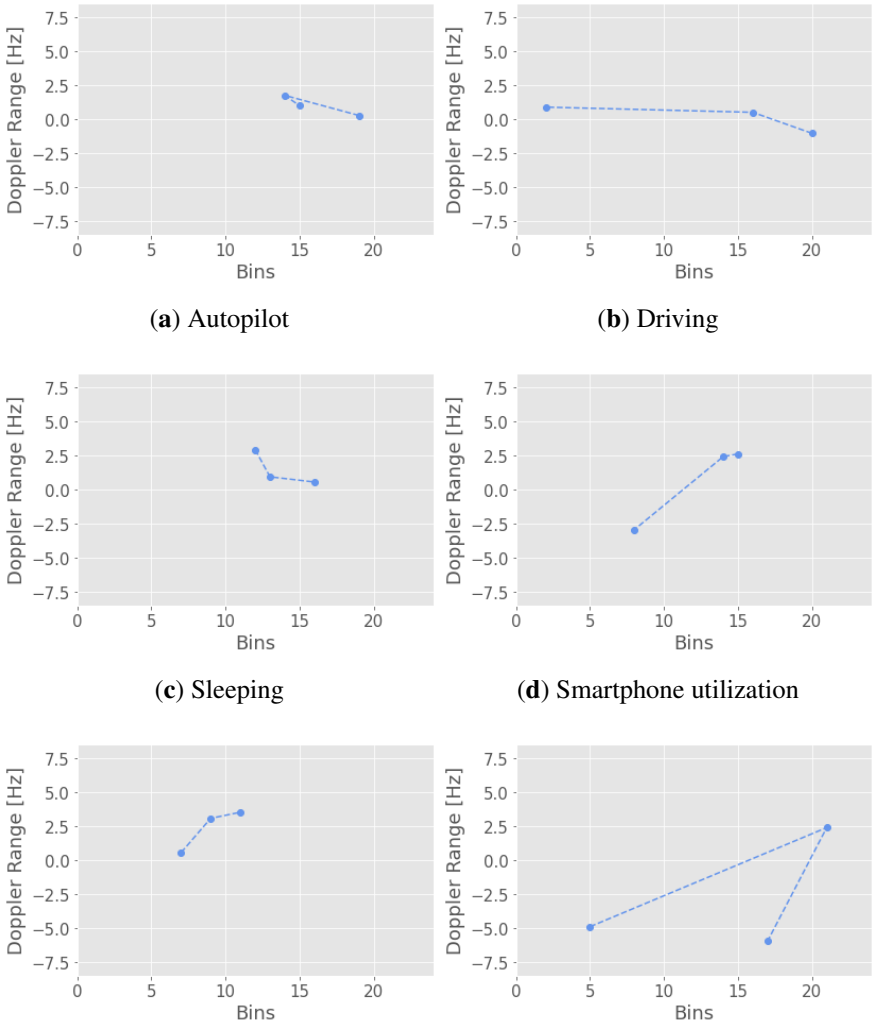
Action	Description
Autopilot	While driving with autopilot, participants were instructed to keep their hands on their knees while sitting in the simulator and observing the virtual environment.
Driving	Participants were asked to drive freely through the virtual city following the general traffic rules. They were also instructed to turn at least once.
Sleeping	For the sleeping action, participants were asked to take a comfortable position in the driving chair while keeping their head in ventral flexion, close their eyes and relax.
Smartphone utilization	The same instruction as for *autopilot* was used, with the addition to check e-mails or social media using their smartphone with both hands.
Driving & Smartphone utilization	During this action, the participants had to perform driving while steering the wheel with the left hand and checking e-mails, social media, etc. using their right hand.
Talking to passenger	A second person was invited as a passenger to take the front seat. The drivers were instructed to actively communicate with the passenger and rotate their head to the passenger while performing regular driving. They could use the right hand for gesticulation.

5.4.2 Experiments

This section describes the performed experiments for driver behavior recognition using range-Doppler maps. The results are reported using Classification Accuracy (correctly classified activity windows divided by the total number of activity windows) and the F1-score for better comparison (see Table 5.6). For the deep learning models, the PyTorch library was selected [161], while for the classic machine learning algorithms, the scikit-learn library was used.

Two different experiments for data splitting and evaluation were run. In the first experiment, random stratified data splitting was used for the acquired radar data into

training (80%) and test (20%) sets as in studies [231, 153, 183, 205]. This was done to evaluate the ability of the architecture to overfit on the radar data of specific persons. In the second experiment, leave-one-out cross-validation was performed , where the whole data of one participant was withheld from the training dataset and used for validation only. The cross-validation was repeated 10 times according to the number of participants. The final accuracy is reported as an average value of over 10 participants (see Figure 5.5). In addition, confusion matrices only for the best-performing models are provided (see Figure 5.8). Importantly, the goal is not to directly compare the classification performance between LSTM and ResNet or LSTM and PVT-Tiny. Given the difference in the model architectures and how they treat data, there is no way to compare them honestly. Instead, rather the performance of these models on the given dataset is evaluated.



(**a**) Autopilot (**b**) Driving

(**c**) Sleeping (**d**) Smartphone utilization

(**e**) Driving & Smartphone utilization (**f**) Talking to passenger

Figure 5.4: Range-Doppler trajectories of six (**a**–**f**) in-cabin activities calculated using the method of [66]. Each trajectory contains a single frame (0.34 s).

5.4.3 Baseline Classification

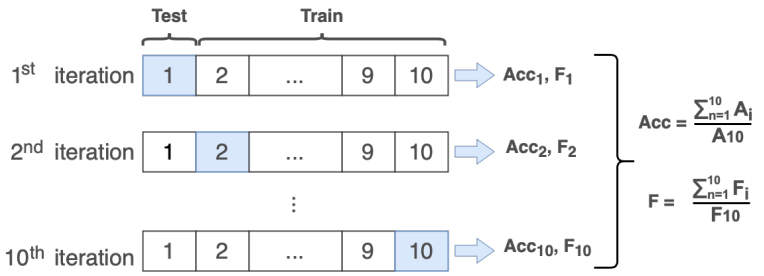


Figure 5.5: Schematic representation of leave-one-participant-out cross-validation method. The number of iterations corresponds to the number of participants in the dataset.

Setting a baseline to compare the proposed method to is challenging due to the very small number of existing radar-based driver monitoring approaches overall, with none providing a source code or a dataset for comparison. Nevertheless, to define a baseline, the method proposed by [66] was re-implemented based on the information provided in their paper. The method generates features by using range-Doppler frames and time-Doppler spectrograms obtained from the in-cabin driver recording. Because of hardware differences, the main focus is put on the features extracted from the range-Doppler trajectory (RDT), in particular, *dynamic Doppler*, *Doppler range* and *dynamic power* because of the similarity to our output data. Among the 12 classifiers evaluated by [66], the ensemble classifier with bagged trees achieved the highest classification accuracy of 93.3% for the range-Doppler trajectory reported on their dataset. The features on the level of single participants using a window size of one second (or three frames) with 2/3 overlap were calculated. A high-pass filter of 10 Hz was not used to mask low-frequency activities and the range of the Doppler was not manipulated as it was proposed in the paper, since this information could be crucial for distinguishing our classes (e.g., hands on the wheel while driving vs. autonomous driving). Instead, the IQR-range normalization was performed where values exceeding the 75^{th} percentile were not considered for the Doppler-trajectory computation. In the next step, following the architecture of the best-performing classifier and the training steps (see [66]), a bagging classifier was built. The training and testing datasets were generated in two ways: splitting the data as equally as possible into ten folds and using nine for training

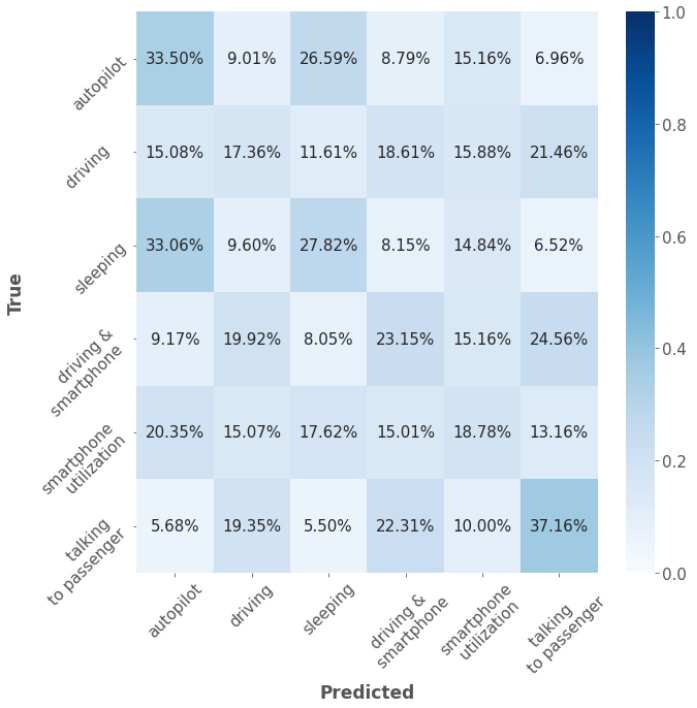
and one for validation as proposed by [66]. Next, leave-one-participant-out cross-validation was performed to achieve a possible comparison to the proposed method (see Figure 5.5). The reported results are the average over the validation splits.

Table 5.5: Baseline classification performance for driving activity recognition on the *RaDA* dataset using re-implementation of Ensemble classifier ([66]*). The obtained results show high similarities for both validation strategies.

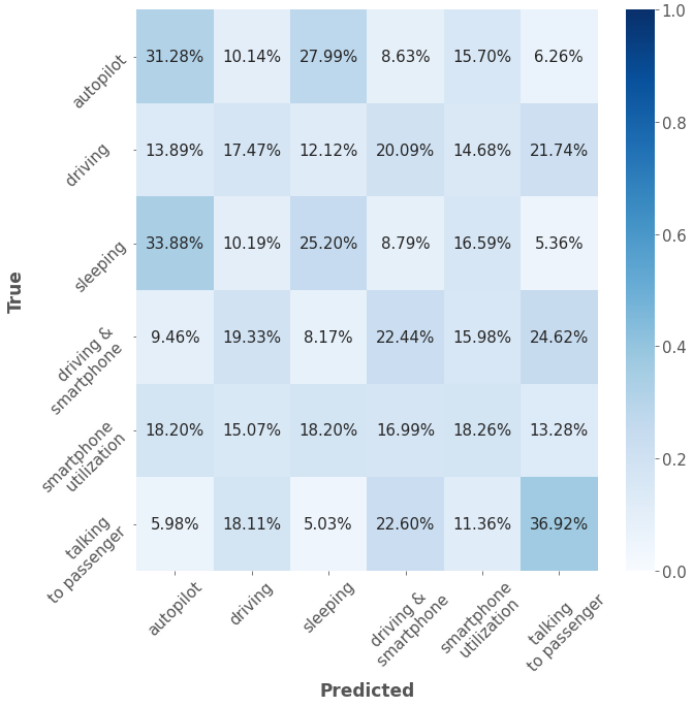
Architecture	Validation Type	IQR Norm.	Accuracy	F1-Score
Ding et al. [66]*	10-fold CV	✓	0.263	0.261
Ding et al. [66]*	LOO CV	✓	0.252	0.249

Table 5.5 and Figures 5.6 (a), (b) represent the obtained results for the classification performance using the ensemble classifier of [66]. Using random stratified data splitting for training and testing, the model achieved an average classification accuracy of 26.3% over six classes. The highest classification accuracy of 37.16% was observed for the class *talking to passenger*. *Autopilot* was the second best-predicted class with an accuracy of 33.50%. The classification accuracy for the four remaining classes was from 0.69 to 11.15 percentage points over the level of random guessing at 16.67%.

After applying leave-one-participant-out cross-validation, a slightly lower classification accuracy but a similar classification pattern was observed. The highest classification accuracy of 36.92% was observed for the class *talking to passenger*, followed by class *smartphone utilization*. The residual classes were either slightly over or under random guessing. The obtained performance drastically deviates from the one reported in the original work of [66]. The low classification performance on the *RaDA* dataset can be explained in several ways. First, fundamental differences between the used UWB radar and the FMCW radar used in the original study. Secondly, the higher sampling rate used in [66] could bear more available data for model training. In the present work, the minimal size of the window was constrained by the frame rate of the used radar. Next, the proposed method did not explicitly consider possible outliers in the data while focusing on the high-frequency components. Finally, in the original work, the rotation, nodding and flexion of the head were the class-discriminating activities, while our data also includes scattering information from the torso. Taking these results together, the proposed method of [66] did not perform well on the *RaDa* data.



(**a**) RDT-features with random stratified data splitting.



(**b**) RDT-features with LOO cross-validation.

Figure 5.6: Confusion matrices of obtained classification results using Ensemble classifier proposed by [66] using random stratified data splitting and leave-one participant-out cross-validation method.

5.4.4 ResNet

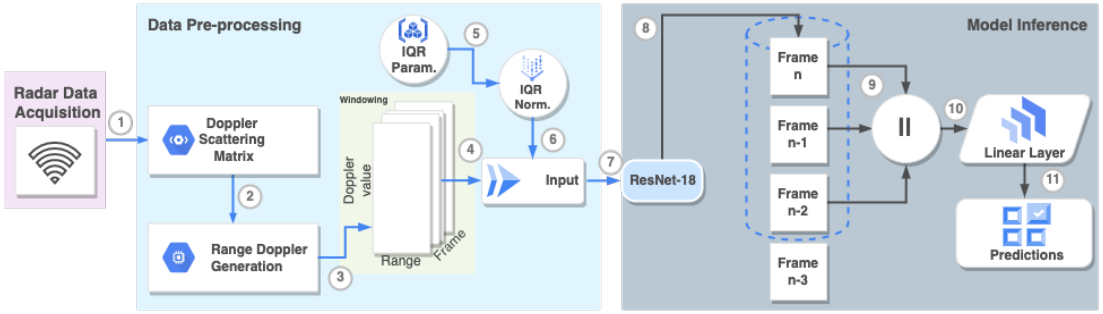


Figure 5.7: Flow diagram of the inference pipeline of the proposed approach. n represents the frame counter. The Doppler data are fed to the ring buffer frame-wise, where the total number of frames capturing one second are concatenated. The concatenated frame data is further forwarded to the linear layer of the network.

The architecture of the proposed ResNet-based approach (see Figure 5.7) was designed with the real-time application in mind. The radar data were transformed into spectrograms which represent range-Doppler maps. Three of them cover approximately a time span of 1 second (more details in Section 5.3.1). Each range-Doppler map is processed independently by the same ResNet-18 to extract features. The features of the last three frames are kept in a ring buffer. This way, whenever a new frame arrives, only this single frame must be processed by ResNet. Then the features of the three frames are concatenated and classified jointly by a fully-connected layer. Training of this architecture was performed using three parallel ResNet-18 instances that share their weights. This ensures the proper flow of the gradients during training and enables training with random shuffling.

For the proposed ring-buffer approach, a PyTorch implementation of ResNet-18 was trained using weights pre-trained on ImageNet-1K. The stochastic gradient descent (SGD) optimizer with a momentum of 0.9 was used. The One Cycle Learning Rate scheduler [197] was selected to decrease the training time. This method is based on the phenomena of 'super-convergence', which can be observed when training with the one-cycle learning rate schedule. Furthermore, the larger possible maximum learning rate can result in an additional increase in classification performance. The maximum

learning rate was set to 0.01. The initial learning rate was chosen to be one-tenth of the maximum learning rate. A mini-batch size of 40 was used and trained for 20 epochs. A higher number of epochs did not lead to any significant improvement in classification performance. For the training and validation, the input data was repeated 3 times in the channel dimension and resized to 224 x 224 pixels.

5.4.5 LSTM

A uni-directional LSTM model was built. The number of features in the hidden state was set to 6, and the number of recurrent layers was 2. The learning rate of 0.001 was used. A dropout layer with a 20% dropout probability was used to prevent overfitting. The mini-batch size was set to 8, and the number of epochs to 80. The number of input features was set to *1024 x 24* corresponding to the Doppler frequency range and bin range. Each action sequence was split into single frames (around 0.34 seconds per frame) for training and validation. Because of a slight variation in the length of obtained recordings, all sequences were cropped to the shortest length of 163 (56.21 s) frames for the model training and evaluation. Data exceeding this range was neglected.

5.4.6 PVT-Tiny

A PVT-Tiny model with pre-trained weights PVTv2-B2 on ImageNet-1k was selected because of it's similarity to the ResNet-18 in terms of model complexity [226]. This enhances the comparison of the performance of these two models on the RaDa dataset. The Adam optimizer was used. To reduce the training time, a Cosine scheduler was deployed. The maximum learning rate was set to 0.001. The mini-batch size was set to 60 and trained for 40 epochs. Decreasing learning rate or increasing the mini-batch size did not lead to any compelling improvements on the average level. The input data was resized to 224 x 224 pixels for model training and validation.

5.4.7 Results

The classification performance for the ResNet-18 architecture is reported in Table 5.6. The perfect average accuracy of 100% over all six classes was achieved with random data splitting with 80% / 20 % ratio both with and without IQR normalization. However,

Table 5.6: Average classification performance for driving activity recognition on the *RaDA* dataset using ResNet-18, LSTM & PVT-Tiny. Due to possible class imbalances, a weighted F1-Score is reported. The highest classification accuracy was achieved by PVT-Tiny without IQR normalization.

Architecture	Validation Type	IQR Norm.	Accuracy	F1-Score
ResNet-18	Random splitting	–	1.0	1.0
ResNet-18	Random splitting	✓	1.0	1.0
ResNet-18	LOO CV	–	0.654	0.621
ResNet-18	LOO CV	✓	**0.714**	**0.687**
LSTM	LOO CV	–	0.439	0.351
LSTM	LOO CV	✓	**0.672**	**0.590**
PVT-Tiny	LOO CV	✓	0.723	0.705
PVT-Tiny	LOO CV	–	**0.749**	**0.729**

a drop in accuracy of 28.6 percentage points was observed for the same architecture when using leave-one-out cross-validation. Without IQR normalization, this decrease was almost 34.6 percentage points. Clearly, the random splitting leads the model to overfit strongly, which is possibly due to the prevalence of features specific to individual persons. In contrast, the models show relatively moderate results when being evaluated using cross-validation. This demonstrates the challenge of inter-person generalization of systems trained with radar data and also the challenge level of the driving monitoring application. Therefore, the models with random splitting are not further considered.

The highest classification accuracy of 71.4% was obtained for the ResNet-18 model using IQR normalization (see Figure 5.8 (a)). The class *autopilot* belongs to the most well-predicted classes with 89.32% accuracy, followed by *smartphone utilization* with 88.34% and *talking to passenger* with 77.41%, respectively. The lowest accuracy values were observed for the classes *driving* and *sleeping* with 56.21% and 57.30%, respectively. Class *driving* had a high confusion with the class *driving & smartphone utilization*, whereas the latter had a high confusion with the classes *driving* and *talking to passenger*. Importantly, all of these three classes shared the same basic driving activity. The position of the right hand and the smartphone utilization's intensity were moderated by the need to maintain the proper lane and avoid any collision, which could impede the class prediction. Similarly, the confusion between the classes *sleeping* and

autopilot can be explained. In the experimental condition, sleeping was defined as a specific head flexion, which depth varied among participants. Considering that in both classes the subject remained still in the driving chair, this confusion rate could be due to the definition of the experimental class. The absence of IQR normalization decreased the average classification accuracy to 65.4%.

For the training and evaluation of the LSTM model, only the leave-one-out cross-validation method was used. The highest average classification accuracy of 67.2% was observed using IQR normalization. The classes *talking to passenger*, followed by *smartphone utilization* achieved the highest classification accuracy (see Figure 5.8) with 96.50% and 94.97% respectively, followed by the class *driving* with 69.54% accuracy. The lowest classification accuracy was observed for the class *driving & smartphone utilization* with 35.98%. The confusion pattern between the classes *sleeping* and *autopilot*; and *driving & smartphone utilization*, *talking to passenger* and *driving* resembled those in the ResNet-18 model. The high confusion between the classes *driving & smartphone utilization* with the classes *driving* and *driving and talking to passenger* can be explained analogously as for ResNet-18. Importantly, the LSTM model received the whole one-minute sequence to estimate a single class. Therefore, the proposed results are for a general model evaluation, not a real-time driver monitoring scenario. The absence of the IQR normalization led to a drop in the classification accuracy to 43.9%. Interestingly, while in the case of ResNet-18, IQR normalization led to an increase of 3.9 percentage points in classification accuracy, for the LSTM model, the difference amounted to 23.3 percentage points.

Unlike LSTM and ResNet-18, PVT-Tiny's highest classification accuracy of 74.9% was observed without IQR normalization. Similarly to the ResNet-18, to the best-predicted classes belong *autopilot*, *smartphone utilization* and *talking to passenger* with 92.24%, 86.38% and 85.06% accuracy, respectively. As in the case of LSTM, for the class *driving and smartphone utilization*, the lowest accuracy of 52.93% was observed. The high confusion between the classes *sleeping* and *autopilot*, as well as *driving and smartphone utilization* and *driving* can be explained in the same manner as for ResNet-18 and LSTM. IQR normalization led to a drop of 2.6 percentage points in classification accuracy (see Table 5.6). This is the most negligible observed difference in the classification accuracy between the presence and absence of IQR normalization compared to ResNet and LSTM.

5.5 Data Augmentation

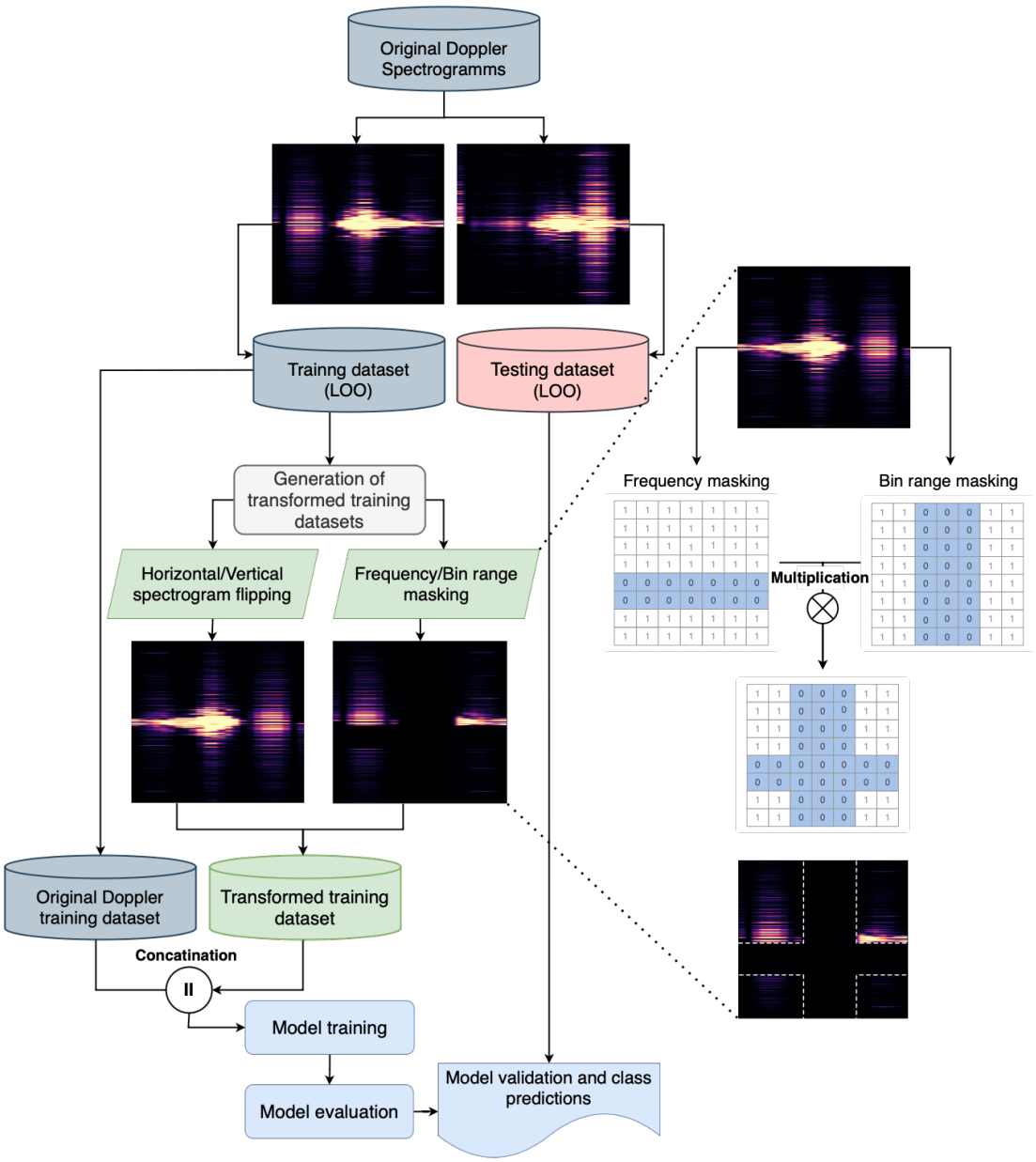


Figure 5.9: Data Augmentation and model training flow diagram. After splitting the data following the leave-one-participant-out policy, a new transformed training data set is generated using flipping masking or frequency/bin range masking strategy. The training is then performed using original data combined with the new, transformed dataset. The model validation is performed on the data of a single, withheld participant.

Section 5.4.7 demonstrated the general feasibility of deep learning algorithms for recognising driving activity associated with engaged, autonomous and distracted driving using the Doppler radar data. The classification performance of the three Deep Learning models significantly outperformed the proposed Machine Learning approach proposed by Ding et al. [66] (see Section 5.4.2). However, in all three deployed Deep Learning models, especially the performance within the classes *driving*, *sleeping* and *driving & smartphone utilization*, requires further improvement. For this purpose, *repeated augmentation* method was employed to extend the number of training samples and thus leverage the models' classification performance. Only the models ResNet and PVT-Tiny are further considered since they demonstrated the highest classification performance.

Regarding the transformation of Doppler spectrograms, two state-of-the-art filters can be deployed: 1) *flipping filter* and 2) frequency and bin range masking using *SpecAugment*. Taylor et al. used horizontal flip and blur filters to generate an augmented radar dataset for activity classification [205]. The deployment of the data augmentation considerably improved the performance of their CNN compared to the model training without any data augmentation. SpecAugment is an augmentation scheme coming from the field of speech recognition [159]. The policy of SpecAugment consists of time masking, frequency masking, and time warping. The authors of this method pointed out a considerable increase in word error rate (WER) when deploying SpecAugment on various speech datasets. Applying a mask for frequency and time domain should increase the variance in training data and contribute to the model's classification performance. Given that those spectrograms in the RaDA data have a bin range instead of the time domain, only frequency and bin masking are deployed, where the latter was performed mimicking the policy of time domain from SpecAugment. The dataset with horizontally and vertically flipped spectrograms was generated by mirroring the entire spectrogram plane along the related axis. Frequency and bin range masking were generated in line with the SpecAugment scheme [159]:

1. Frequency masking: The size m for the mask size was chosen from a uniform distribution from 0 to M, where the maximum value of M was set to 40% of the total frequency range size. The Doppler-frequency values $[m_0,\ m_0 + m)$ were then masked, where m_0 is selected from $[0, f - m]$, f denotes the frequency dimension of the spectrogram.

2. Bin range masking: The size b for the mask size was chosen from a uniform distribution from 0 to B, where the maximum value of B was set to 40% of the total frequency range size. The bins within the range $[b_0,\ b_0 + b)$ were then masked, where b_0 is selected from $[0,\ r - m]$, r denotes the bin dimension of the spectrogram.

The resulting mask could cover the frequency and bin range axes (see Figure 5.9) or one specific axis. Subsequently, the obtained mask was employed across all three spectrograms representing one second (see Figure 5.3).

Figure 5.9 shows the flow diagram with procedural steps for building augmented datasets, subsequent model training and evaluation. Two novel datasets *Horizontal/Vertical flipping* and *Frequency/Bin range masking* were built upon the training dataset after separating it from the testing dataset following the leave-one participant out schema. Next, the novel dataset was concatenated with the original dataset, and the resulting augmented dataset was used for the model training. The data augmentation was performed depending on the impact of the IQR-normalization method on a particular model from Section 5.4.7. Thus, the data augmentation for the ResNet was performed using IQR-normalization, whereas the data augmentation for the PVT-Tiny was built upon the raw data. The final model validation and performance score estimations were performed on a withheld participant's testing dataset using a leave-one-out scheme.

Table 5.7: Average classification performance of ResNet-18 and PVT-Tiny for driving activity recognition on the *RaDA* dataset using Flipping and Frequency/Bin range masking techniques compared to raw input and IQR transformed data. Results in bold represent improvemements against the baseline.

Class	**ResNet-18**				**PVT-Tiny**			
	RAW	**IQR***	**Flip**	**Mask**	**RAW***	**IQR**	**Flip**	**Mask**
Autopilot	0.651	0.800	0.818	**0.820**	0.838	0.866	0.810	0.802
Driving	0.537	0.555	0.523	0.531	0.614	0.525	0.641	**0.656**
Sleeping	0.549	0.558	**0.600**	0.559	0.681	0.677	0.656	0.654
Smartphone utilization	0.802	0.893	0.895	**0.860**	0.865	0.883	0.865	**0.867**
Driving & **Smartphone utilization**	0.534	0.543	**0.555**	0.525	0.573	0.571	0.606	**0.610**
Talking to passenger	0.632	0.752	**0.774**	0.754	0.785	0.687	**0.775**	0.763

Raw: Absence of IQR transformation. *: Model-specific baseline.

5.5.1 Results

The classification performance of ResNet-18 and PVT-Tiny using an augmented dataset are summarized in Figure 5.10 and Table 5.7. The results are further discussed, focusing on weighted F1-score since it can better account for multiclass and class imbalance (see Section 2.4).

The highest average classification performance for ResNet-18 was obtained when applying IQR transformation, and this result is further considered in this section as a baseline. Training the ResNet-18 model with the second dataset containing horizontal and vertical flip filters led to the model improvement at 0.6 percentage points compared to the baseline (see Figure 5.10). By detailed examination of the class-specific F1-score (see Table 5.7), it can be seen that for the class *sleeping*, the classification performance increased at 4.2 percentage points, achieving an average F1-score of 60 %. For the class *autopilot*, an average improvement of 1.8 percentage points was observed. Finally, there was an increase in the average F1-score for the class *Driving & Smartphone utilization*. This comparably low average improvement (see [205]) suggests that individual variance among subjects can not be accounted for simply by augmenting data with image-flipping techniques. This is is partly proved by the individual classification results as shown in Figure A1.

In contrast, adding the second dataset with masked frequency/bin range axes to model training led to a drop in classification performance at 1.1% point. Apart from the classes *sleeping* with an average improvement of 2% point, the performance was either neglectable better or below the baseline (see Table 5.7). One possible explanation for this decrease in classification performance can be too high information loss in the spectrogram through the frequency or time masking. Thus, decreasing the mask range on x and y axes or adapting the mask to the network architecture can enhance classification performance. Next, ResNet might encounter difficulties handling masked spectrograms due to its reliance on spatial information. Frequency/bin range masking might introduce potential disruption in spatial relationships between neighbour regions in a spectrogram and the spatial locality of the features. Tuning the hyperparameters related to the filter and stride size might enhance the classification performance when training the model using masked data. Finally, the max pooling operation used in the default ResNet-18 architecture can contribute to difficulty handling masked data. The

network may lose crucial information required to understand the input data by omitting masked regions through the max pooling operation. To mitigate this information loss, alternative pooling strategies should be evaluated.

None of the two data augmentation techniques yielded any improvements in the classification performance of the PVT-Tiny model compared to its baseline (see Figure 5.10). Since spectrogram flipping and spectrogram masking provide a very different representation of the input data, it is logical to assume that the core reason lies in the feature extraction and feature learning policies of PVT-Tiny. Presumably, PVT-Tiny may not have been properly fine-tuned for training with the augmented data. Next, the augmented dataset could expose unrealistic or irrelevant variations or noise in the data, leading to overfitting. This might result in low generalization power of the model to unseen data causing degraded performance. Next, In contrast to CNNs which possess the ability of translational invariance (see Section 2.3.5), PVT-Tiny relies on self-attention mechanisms which allow it to capture global dependencies among different regions of an image (or a spectrogram). Flipping the spectrogram along the horizontal and vertical axes can potentially alter the relationship among different image regions diminishing the self-attention mechanism's ability to learn patterns and features relevant to correct class prediction. Consequently, image flipping on horizontal and vertical axes might harm the model performance. Finally, SpecAugment might generate image modifications violating assumptions of ResNet and PVT-Tiny as the masked regions may not conform to the expected local and global spatial patterns the models are trained to recognize.

Figures A1 and A2 additionally provide individual F1-scores by the model type and performed transformation. It can be clearly seen that for all participants, at least one out of six classes was predicted with a F1-score of over 80 %. Frequently, while improving the prediction of one class through data augmentation, the performance in other classes drops. This could indicate model overfitting, which occurred due to additional variance gained through the data augmentation. On the other hand, the selected data augmentation techniques could introduce new feature distributions which were not representative for the true data. Thus, the models could struggle to classify the data correctly.

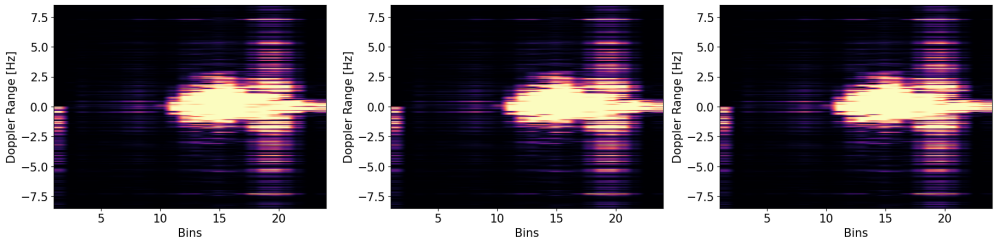
5.6 Conclusion

This chapter introduces a novel method for recognizing normal (engaged), autonomous or distracted driving behavior using a UWB Radar and deep neural networks. The proposed approach has demonstrated the sufficiency of Doppler-frequency information combined with deep neural networks for the recognition of six different driving activities. Moreover, the state-of-the-art machine learning method proposed by Ding et al. [66] that uses Doppler-trajectory features was outperformed by ResNet-18, LSTM and PVT-Tiny as Section 5.4.7 demonstrates. Furthermore, different cross-validation techniques were evaluated. While the frequently used stratified method for data splitting achieved the highest possible accuracy, a more strict leave-one participant-out cross-validation demonstrated that radar-based activity recognition could not be easily generalized to a new, unseen driver. Another novelty of this work is deploying a Transformer network PVT-Tiny which achieved the highest average classification accuracy. Next, a universal interquartile range (IQR) normalization method was applied to the Doppler data to resolve multiple pre-processing steps that might depend on a particular radar system. It additionally significantly boosted the classification accuracy of ResNet-18 and LSTM but not those of PVT-Tiny. IQR-normalization could exclude the high-frequency related values carrying class-discriminating information crucial for PVT-Tiny. The simpler architectures of ResNet-18 and LSTM probably did not allow the creation of more complex representations of the classes when IQR-normalization was disabled.

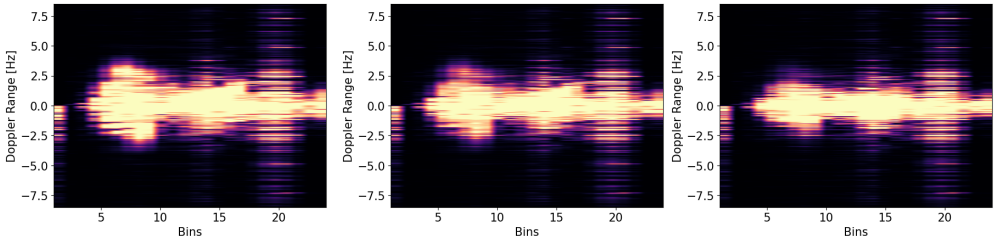
Despite the overall good classification performance, the classification accuracy within the classes *driving*, *sleeping* and *driving & smartphone utilization* was significantly below the average accuracy value. To improve the model accuracy, especially for these classes, in Section 5.5, two additional datasets were generated, where the original data were either flipped along horizontal and vertical axes or masked along frequency or bin range axes using the policy of SpecAugment method. For ResNet-18, the model training with additional vertical and horizontal spectrogram flipping resulted in a minor improvement, although the F1-score for the class *sleeping* was drastically increased. For PVT-Tiny, there was a drop in model classification performance when deploying augmented data for training. As the results showed, SpecAugment as an augmentation method for Doppler-range data proved false for both Resnet and PVT-Tiny. It can be concluded that the deployed data augmentation techniques could not account for the

high variance in driving activity patterns among participants. The high confusion among the classes *driving & smartphone utilization* with the classes *driving* and *driving and talking to passenger*, as well as *autopilot* and *sleeping* can be explained through similar and partly analogue driving behavior pattern presented in these classes.

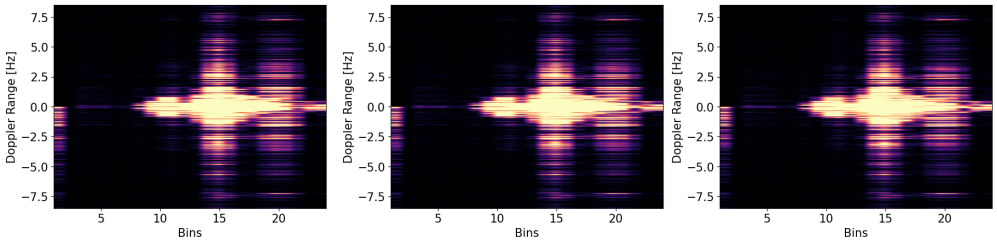
In future work, a large-scale evaluation of the proposed approach under varying driving conditions and a larger number of participants could be performed. Next, it is also important to investigate ways to reduce the ambiguity between the classes with high confusion levels, as it is vital for the correct system response in safety-critical driving scenarios. It can be achieved while deploying novel data annotation strategies, for example, a moment-to-moment annotation combined with a video camera. It is also essential to evaluate the impact of various window sizes of the Doppler range (for example, 0.5 seconds or 2 seconds) on the model classification performance. Finally, additional augmentation techniques, such as noise injection or more advanced signal processing techniques, can be explored to improve the performance and robustness of the deep learning models and make them more effective in handling the data of new, unseen drivers.



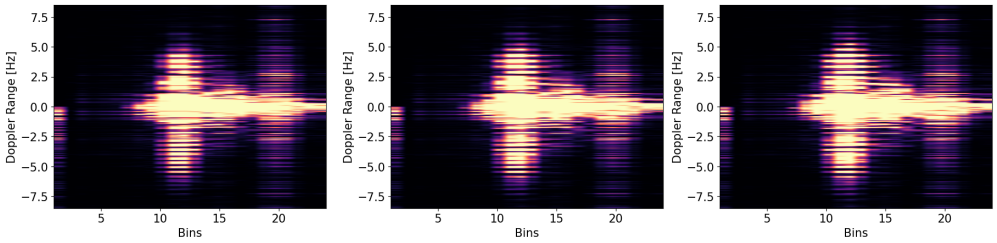
(**a**) Autopilot



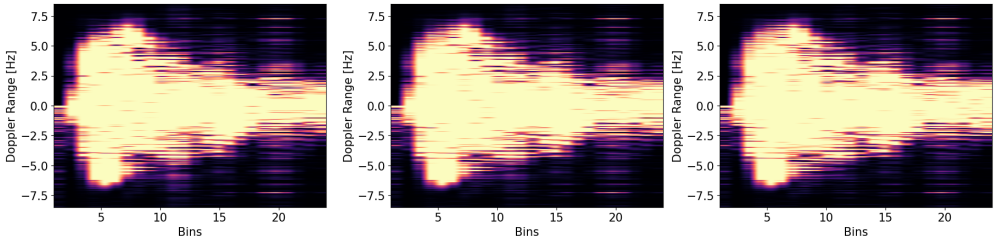
(**b**) Driving



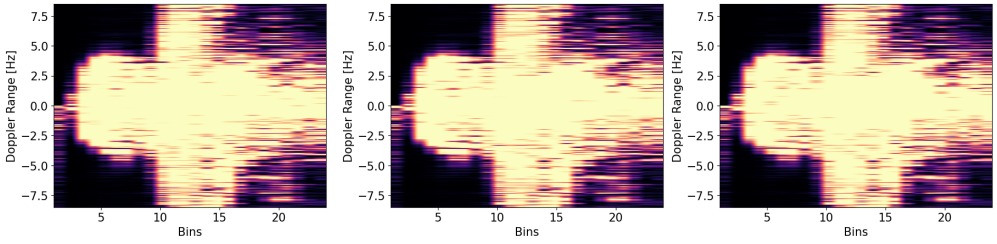
(**c**) Sleeping



(**d**) Smartphone utilization

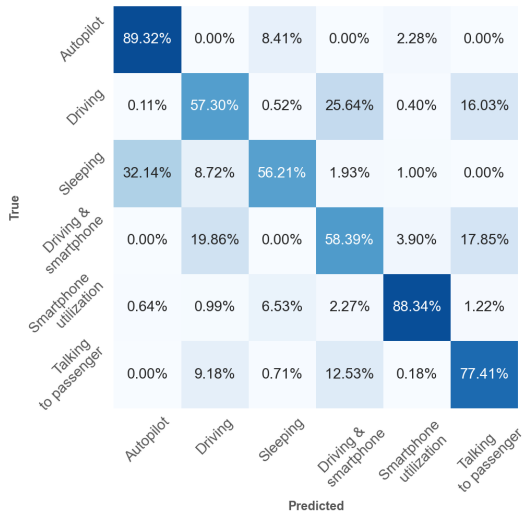


(**e**) Driving & Smartphone Utilization

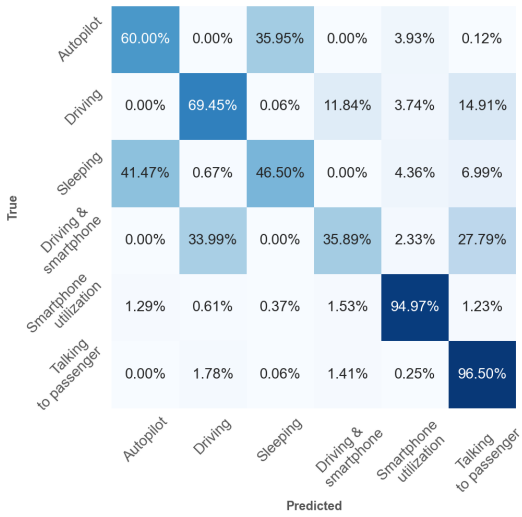


(**f**) Talking to passenger

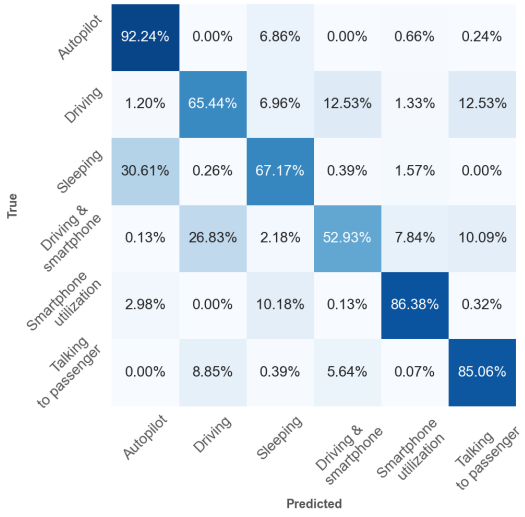
Figure 5.3: Range-Doppler spectrograms of six (**a**–**f**) in-cabin activities captured by the radar. Three images within one class represent roughly one second.



(**a**) Confusion matrix for Resnet-18 with LOO cross-validation.

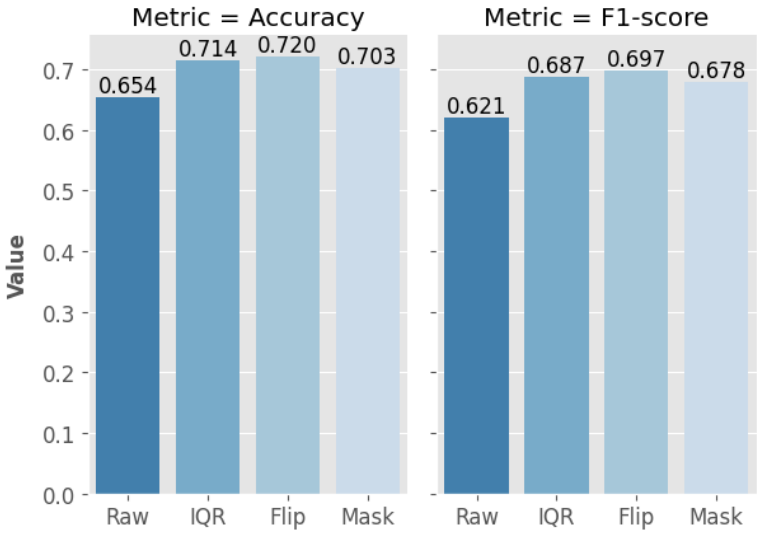


(**b**) Confusion matrix for LSTM with LOO cross-validation.

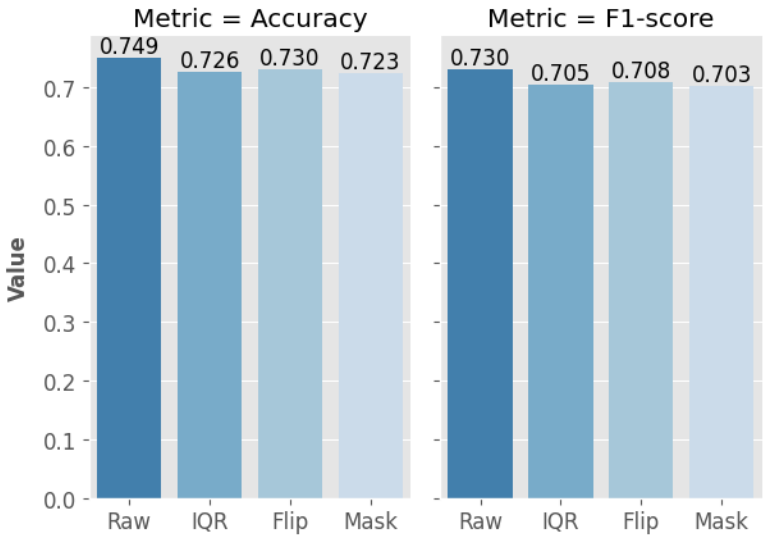


(**c**) Confusion matrix for PVT-Tiny with LOO cross-validation.

Figure 5.8: Confusion matrices of the best classification results using ResNet-18, LSTM and PVT-Tiny.



(**a**) ResNet-18



(**b**) PVT-Tiny

Figure 5.10: Average accuracy and F1-Score by data augmentation and transformation techniques for ResNet-18 and PVT-Tiny. For ResNet-18, adding the second one with horizontal/vertical flipping to the training dataset resulted in a slight model improvement.

6 Conclusion

This thesis investigated the feasibility of physiological sensors combined with machine learning algorithms for the potential development of mental imagery-aware systems. Considering the versatility of forms and functions of mental imagery, its two most common forms were taken as a core of mental imagery-aware systems, namely mind-wandering and spatial imagery. The latter was derived from the user's ongoing engagement. Thereby, two scenarios for the potential mental-imagery aware systems were explored: (1) quantification of mind-wandering in a learning scenario using electrodermal and gaze features, and (2) spatial imagery or engagement recognition using gaze features and radar data in a driving condition. The primary goal was to pave the way for scalable solutions compatible with applications and systems accompanying daily activities. The work summary is provided in Section 6.1, and future work is suggested in Section 6.2.

6.1 Summary

Chapter 1 proposed the central research question of this thesis, namely **"Can machine learning combined with physiological sensors enhance the quantification of mental imagery in the context-ware systems?"**. This research question was motivated by the fact that despite continuously increasing automation levels of the systems and their ability to track users' physical activity, the mental dimension has yet to be considered in context-aware systems.

How can the features extracted from the electrodermal activity and eye movement, along with machine learning, contribute to detecting episodes of mind wandering in a learning context?

Chapter 3 introduced a novel approach for the recognition of the episodes of mind-wandering in the learning setting using electrodermal features and gaze data. By extensive statistical analysis of the experienced thought type and related environmental conditions, as introduced in Section 3.4, it was shown that mind-wandering should not necessarily always have a detrimental impact on the learning performance but might be required for proper information comprehension. In Section 3.5.4, the feasibility of using an EDA sensor as a single device for mind-wandering detection was demonstrated for the first time. It was also shown that combining EDA with eye-tracking features achieves the highest classification accuracy for mind-wandering quantification. The highest classification accuracy for all the sensor types was observed by using the Random Forest model. Thus, the combination of EDA features with machine learning algorithms can serve as a strong base for mind-wandering-aware systems.

How can machine learning contribute to detecting spatial imagery and engagement with a driving task under various driving conditions?

Chapter 4 elaborates an extensive study to investigate the impact of autonomous versus manual driving on the driver's spatial imagery. This work was motivated by the existing gap in understanding of long-term impacts of autonomous driving on human spatial imagery. Section 4.5 demonstrated compelling results in structural differences in cognitive workload, mental maps, and gaze movement patterns depending on the driving modality. These finds were taken as a core of the study investigating spatial imagery-aware systems in a driving setting. Thus, it was shown that eye-tracking technology could classify active navigation or the high engagement of the driver versus a passive transportation mode, as proposed in Section 4.6. The experimental evaluation demonstrated the highest classification accuracy using the Gradient Boosting algorithm that outperformed the state-of-the-art in the comparable driving setting. Considering the observed results, this chapter offers a prospect for a driver's engagement-aware system.

How can the radar system contribute to driver-independent, privacy-driven monitoring solutions in the context of engagement-aware systems?

Chapter 5.1 introduces a quasi-optical solution for drivers' engagement-aware system. For this purpose, a UWB pulse-Doppler radar was used to record the driver's activity associated with attentive, autonomous and distracted driving behavior. Overall, the chapter demonstrated the great potential of the radar system to discriminate among different driving behaviors in a driver-independent fashion.

Section 5.4.3 demonstrated the low performance of the state-of-the-art method consisting of Doppler-trajectory-related features and Ensemble classifier on the obtained radar data. In contrast, Section 5.4.7 demonstrated that CNN, LSTM and a Pyramid Visual Transformer trained on Doppler range data can achieve a reasonable user-independent classification accuracy. Further, it was demonstrated that the IQR normalization method could enhance the generalization ability of CNN and LSTM, resolving the necessity of extensive pre-processing steps for radar data.

In Section 5.5 two data augmentation techniques were additionally evaluated to account for the repeatedly occurring confusion among particular classes in ResNet-18 and Pvt-Tiny. The results demonstrated that conventional vertical and horizontal flipping filters and SpecAugment-based frequency and bin range masking did not make any compelling improvements in the classification performance, showing even an opposite effect in some experimental outcomes. The work suggests that these two particular augmentations techniques can not attribute the observed individual variance in performed activities but instead bear a ground for model overfitting. Taken together, there is a great potential for the radar combined with deep learning algorithms in the context of engagement-aware systems.

6.2 Limitations & Future Work

The proposed work reveals the potential of physiological, low-cost sensors and machine learning algorithms for mental-imagery-aware systems. Thus, many contributions of this work can be considered as building blocks towards mental imagery-aware systems. Nevertheless, there is still room for improvement.

Data-annotation validity. The proposed method for mind-wandering quantification in a learning scenario requires at least 30 seconds of the input signal to detect an episode of mind-wandering. Thus, the system is not sensitive to the mind-wandering related signal changes within shorter time windows. Next, the labelling of the episodes of mind wandering itself does not provide a sufficient ground truth level since it strongly relates to the self-awareness ability of the users.

Labels used for radar data also experienced a fluctuating validity. The selected annotation strategy considered the entire recording per driving activity belonging to a single class. However, this could be deceptive for some classes with interfering or similar activities (e.g. driving versus driving and smartphone utilization; driving with autopilot versus sleeping). Thus, a moment-to-moment data annotation strategy should be considered in future work.

Data augmentation and Model training. Despite the proposed radar-based engagement recognition demonstrating overall feasibility for future radar-based engagement-aware systems, further improvements in terms of classification accuracy are of vital importance. Considering the high variation among drivers, including more drivers in the RaDa dataset is necessary. It is also reasonable to deploy more sophisticated data augmentations strategies that can produce more variations in the data representation. Generative Adversarial Networks could be a first step to improve the classification performance of the models upon already existing data.

A more dedicated fine-tuning of hyperparameters, for example, changing the filter and stride size of ResNet-18, can improve the classification performance on the existing RaDa dataset. Also, the transfer learning approach, where a model pre-training on a similar radar dataset takes place before starting with RaDa, might enhance the classification performance on the RaDa dataset.

Driving environment. Considering driver's engagement-aware systems, the main limitation for both radar- and eye-tracking-based systems is that the studies were performed in the driving simulator. A further validation study under real driving conditions is strongly required to prove the assumptions and classification results observed under simulated driving conditions.

Appendix

Figure A1: Individual F1-scores by driving class and augmentation/normalization technique obtained by ResNet-18.

Figure A2: Individual F1-scores by driving class and augmentation/normalization technique obtained by PVT-Tiny.

Bibliography

[1] E4 wristband. *https://www.empatica.com/research/e4/*, 2020. *9, 39, 61*

[2] Tobii eye tracker 4c. *https://gaming.tobii.com/tobii-eye-tracker-4c/*, 2020. *61*

[3] Apple watch user guide. *https://support.apple.com/guide/watch/manage-fall-detection-apd34c409704/watchos*, 2022. *9, 20*

[4] Convolutional neural networks cheatsheet. *https://stanford.edu/cheatsheet-convolutional-neural-networks*, 2022. *48*

[5] X4m200 datasheet. *novelda_x4m200_respiration_sensor_rev_b_preliminary_2.pdf*, 2022. *118*

[6] Abowd, Gregory Abowd, Anind Dey, Anind K, Brown, Peter J, Davies, Nigel, Smith, Mark Sandler, Steggles, and Pete. Towards a better understanding of context and context-awareness. In *International symposium on handheld and ubiquitous computing*, pages 304–307. Springer, 1999. *19*

[7] Daron Acemoglu and Pascual Restrepo. Artificial intelligence, automation and work. *SSRN Electronic Journal*, 01 2018. *84*

[8] Abien Fred Agarap. Deep learning using rectified linear units (relu), 2018. *48*

[9] Vlada Aginsky, Catherine Harris, Ronald Rensink, and Jack Beusmans. Two strategies for learning a route in a driving simulator. *Journal of Environmental Psychology*, 17:317–331, 12 1997. *35*

[10] Negar Ahmadpoor and Sina Shahab. Spatial knowledge acquisition in the process of navigation: A review. *Current Urban Studies*, 7:1–19, 03 2019. *34*

[11] Shahzad Ahmed, Dingyang Wang, Junyoung Park, and Sung Ho Cho. Uwb-gestures, a public dataset of dynamic hand gestures acquired using impulse radar sensors. *Scientific Data*, 8:1–9, 04 2021. *114*

[12] Haldun Akoglu. User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, 18, 08 2018. *100*

[13] Negar Alinaghi, Markus Kattenbeck, Antonia Golab, and Ioannis Giannopoulos. Will you take this turn? gaze-based turning activity recognition during navigation. In *11th International Conference on Geographic Information Science (GIScience 2021)-Part II*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2021. *37, 102, 103, 105, 111*

[14] Gary L. Allen, editor. *Human spatial memory*. Erlbaum, Mahwah, N.J., 2004. *34*

[15] Donald Appleyard. Styles and method of structuring a city. *Environment and Behavior*, 2:100–117, 06 1970. *84*

[16] Audacity. *https://manual.audacityteam.org/man/dc_offset.html*. *64*

[17] Benjamin Baird, Jonathan Smallwood, and Jonathan W Schooler. Back to the future: Autobiographical planning and the functionality of mind-wandering. *Consciousness and cognition*, 20(4):1604–1611, 2011. *58*

[18] Evelyn Barron Millar, Leigh Riby, Joanna Greer, and Jonathan Smallwood. Absorbed in thought: The effect of mind wandering on the processing of relevant and irrelevant events. *Psychological science*, 22:596–601, 04 2011. *32, 33, 57*

[19] Mathias Benedek and Christian Kaernbach. A continuous measure of phasic electrodermal activity. *Journal of neuroscience methods*, 190(1):80–91, 2010. *39, 40*

[20] Mathias Benedek, Robert Stoiser, Sonja Walcher, and Christof Körner. Eye behavior associated with internally versus externally directed cognition. *Frontiers in Psychology*, 8:1092, 2017. *59, 78*

[21] Candice Bentéjac, Anna Csörgő, and Gonzalo Martínez-Muñoz. A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, 54(3):1937–1967, 2021. *45, 46*

[22] Robert Bixler and Sidney D'Mello. Automatic gaze-based detection of mind wandering with metacognitive awareness. In *User Modeling, Adaptation and Personalization: 23rd International Conference, UMAP 2015, Dublin, Ireland, June 29–July 3, 2015. Proceedings 23*, pages 31–43. Springer, 2015. *32, 59, 64*

[23] Robert Bixler and Sidney D'Mello. Automatic gaze-based user-independent detection of mind wandering during computerized reading. *User Modeling and User-Adapted Interaction*, 26(1):33–68, 2016. *32, 33, 57, 60, 75*

[24] Nathaniel Blanchard, Robert Bixler, Tera Joyce, and Sidney D'Mello. Automated physiological-based detection of mind wandering during learning. In *International conference on intelligent tutoring systems*, pages 55–60. Springer, 2014. *32, 60, 64, 75*

[25] Blender. *https://www.blender.org/*. *92*

[26] Alessia Bocchi, Marika Carrieri, Stefania Lancia, Valentina Quaresima, and Laura Piccardi. The key of the maze: The role of mental imagery and cognitive flexibility in navigational planning. *Neuroscience Letters*, 651:146–150, 2017. *33*

[27] Mohammud J Bocus, Wenda Li, Shelly Vishwakarma, Roget Kou, Chong Tang, Karl Woodbridge, Ian Craddock, Ryan McConville, Raul Santos-Rodriguez, Kevin Chetty, et al. Operanet, a multimodal activity recognition dataset acquired from radio frequency and vision-based sensors. *Scientific data*, 9(1):1–18, 2022. *114*

[28] Bosch. AInterior monitoring systems. How monitroing the vehicle interior increases safety, comfort, and conveniece. *https://www.bosch-mobility-solutions.com/en/solutions/interior/interior-monitoring-systems/*, 2022. *9, 20*

[29] Nigel Bosch and Sidney D'Mello. Automatic detection of mind wandering from video in the lab and in the classroom. *IEEE Transactions on Affective Computing*, pages 1–1, 2019. *33, 57*

[30] Kevin Bouchard, Julien Maitre, Camille Bertuglia, and Sébastien Gaboury. Activity recognition in smart homes using uwb radars. *Procedia Computer Science*, 170:10–17, 2020. The 11th International Conference on Ambient Systems, Networks and Technologies (ANT) / The 3rd International Conference on Emerging Data and Industry 4.0 (EDI40) / Affiliated Workshops. *41*

[31] Claire Braboszcz and Arnaud Delorme. Lost in thoughts: neural markers of low alertness during mind wandering. *Neuroimage*, 54(4):3040–3047, 2011. *32, 33, 57*

[32] Jason J Braithwaite, Derrick G Watson, Robert Jones, and Mickey Rowe. A guide for analysing electrodermal activity (eda) & skin conductance responses (scrs) for psychological experiments. *Psychophysiology*, 49(1):1017–1034, 2013. *40*

[33] Christian Braunagel, Wolfgang Rosenstiel, and Enkelejda Kasneci. Ready for take-over? a new driver assistance system for an automated classification of driver take-over readiness. *IEEE Intelligent Transportation Systems Magazine*, 9(4):10–22, 2017. *103*

[34] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001. *45, 46*

[35] Iuliia Brishtel, Anam Ahmad Khan, Thomas Schmidt, Tilman Dingler, Shoya Ishimaru, and Andreas Dengel. Mind wandering in a multimodal reading setting: Behavior analysis & automatic detection using eye-tracking and an eda sensor. *Sensors*, 20(9):2546, 2020. *10, 58, 63, 73*

[36] Iuliia Brishtel, Stephan Krauss, Mahdi Chamseddine, Jason Raphael Rambach, and Didier Stricker. Driving activity recognition using uwb radar and deep neural networks. *Sensors*, 23(2), 2023. *12, 114, 116*

[37] Iuliia Brishtel, Stephan Krauß, Thomas Schmidt, Jason Raphael Rambach, Igor Vozniak, and Didier Stricker. Classification of manual versus autonomous driving based on machine learning of eye movement patterns. In *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 700–705, 2022. *82, 114*

[38] Iuliia Brishtel, Thomas Schmidt, Igor Vozniak, Jason Raphael Rambach, Bruno Mirbach, and Didier Stricker. To drive or to be driven? The impact of autopilot, navigation system, and printed maps on driver's cognitive workload and spatial knowledge. *ISPRS International Journal of Geo-Information*, 10(10), 2021. *82, 114*

[39] J.D. Bryan, Junehee Kwon, Nathanial Lee, and Yang Kim. Application of ultra-wide band radar for classification of human activities. *Radar, Sonar & Navigation, IET*, 6:172–179, 03 2012. *41*

[40] Gary Burnett. Turn right at the traffic lights: The requirement for landmarks in vehicle navigation systems. *Journal of Navigation*, 53(3):499–510, 2000. *34*

[41] Gary Burnett and Kate Lee. The effect of vehicle navigation systems on the formation of cognitive maps. *Traffic and Transport Psychology: Theory and Application*, 01 2005. *34*

[42] Gary Burnett and Kate Lee. The effect of vehicle navigation systems on the formation of cognitive maps. *Traffic and Transport Psychology: Theory and Application*, 01 2005. *34, 35, 81, 84*

[43] Georg Buscher, Andreas Dengel, and Ludger van Elst. Eye movements as implicit relevance feedback. In *CHI '08 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '08, page 2991–2996. Association for Computing Machinery, New York, NY, USA, 2008. *37, 38*

[44] Christopher Cabrall, Alexander Eriksson, Felix Dreger, Riender Happee, and Joost de Winter. How to keep drivers engaged while supervising driving automation? a literature survey and categorisation of six solution areas. *Theoretical Issues in Ergonomics Science*, 22:332–365, 03 2019. *114*

[45] Yekta Can, Niaz Chalabianloo, Deniz Ekiz, and Cem Ersoy. Continuous stress detection using wearable sensors in real life: Algorithmic programming contest case study. *Sensors*, 19, 04 2019. *74*

[46] Delphine Caruelle, Anders Gustafsson, Poja Shams, and Line Lervik-Olsen. The use of electrodermal activity (eda) measurement to understand consumer

emotions – a literature review and a call for action. *Journal of Business Research*, 104:146–160, 2019. *95*

[47] Gianna Cassidy and Raymond Macdonald. The effects of music choice on task performance: A study of the impact of self-selected and experimenter-selected music on driving game performance and experience. *Musicae Scientiae*, 13(2):357–386, 2009. *59*

[48] Gavin Cawley and Nicola Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11:2079–2107, 07 2010. *75, 104*

[49] Peter R. Chapman and Geoffrey Underwood. Chapter 17 – Visual search of dynamic scenes: Event types and the role of experience in viewing driving situations. In Geoffrey Underwood, editor, *Eye Guidance in Reading and Scene Perception*, pages 369–393. Elsevier Science Ltd, Amsterdam, 1998. *111*

[50] Nitesh Chawla, Kevin Bowyer, Lawrence Hall, and W. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *J. Artif. Intell. Res. (JAIR)*, 16:321–357, 06 2002. *75*

[51] Marcus Cheetham, Cátia Cepeda, and Hugo Gamboa. Automated detection of mind wandering: A mobile application. In *Proceedings of the International Joint Conference on Biomedical Engineering Systems and Technologies*, BIOSTEC 2016, pages 198–205, Portugal, 2016. SCITEPRESS - Science and Technology Publications, Lda. *32, 61*

[52] Siyuan Chen, Julien Epps, Natalie Ruiz, and Fang Chen. Eye activity as a measure of human mental effort in hci. In *Proceedings of the 16th International Conference on Intelligent User Interfaces*, IUI '11, page 315–318, New York, NY, USA, 2011. Association for Computing Machinery. *37, 38*

[53] Minho Choi, Gyogwon Koo, Minseok Seo, and Sang Woo Kim. Wearable device-based system to monitor a driver's stress, fatigue, and drowsiness. *IEEE Transactions on Instrumentation and Measurement*, 67(3):634–645, 2018. *84*

[54] Elizabeth Chrastil and William Warren. Active and passive spatial learning in human navigation: Acquisition of graph knowledge. *Journal of experimental psychology. Learning, memory, and cognition*, 41, 11 2014. *84*

[55] Ding Congzhang, Yong Jia, Guolong Cui, Chuan Chen, Xiaoling Zhong, and Yong Guo. Continuous human activity recognition through parallelism lstm with multi-frequency spectrograms. *Remote Sensing*, 13:4264, 10 2021. *113*

[56] Corinna Cortes and Vladimir Naumovich Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 2004. *44*

[57] Adele Cutler, D Richard Cutler, and John R Stevens. Random forests. In *Ensemble machine learning*, pages 157–175. Springer, 2012. *45, 46*

[58] Hercules Dalianis and Hercules Dalianis. Evaluation metrics and evaluation. *Clinical text mining: secondary use of electronic patient records*, pages 45–53, 2018. *53*

[59] Christine Dancey and John Reidy. *Statistics Without Maths for Psychology (7th edition)*. Pearson, 05 2017. *100*

[60] Joel Dapello, Tiago Marques, Martin Schrimpf, Franziska Geiger, David Cox, and James J DiCarlo. Simulating a primary visual cortex at the front of cnns improves robustness to image perturbations. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 13073–13087. Curran Associates, Inc., 2020. *47*

[61] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240, 2006. *53*

[62] Michael E Dawson, Anne M Schell, and Diane L Filion. The electrodermal system. *Handbook of psychophysiology*, 2:200–223, 2007. *39, 40, 78, 79*

[63] Ankita Dey, Sreeraman Rajan, George Xiao, and Jianping Lu. Fall event detection using vision transformer. In *2022 IEEE Sensors*, pages 1–4. IEEE, 2022. *117*

[64] Ricardo Alfredo Cajo Diaz, Mihaela Ghita, Dana Copot, Isabela Roxana Birs, Cristina Muresan, and Clara Ionescu. Context aware control systems: An engineering applications perspective. *IEEE Access*, 8:215550–215569, 2020. *19*

[65] Murat Dikmen and C. Burns. Autonomous driving in the real world: Experiences with tesla autopilot and summon. *Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 2016. *35*

[66] Chuanwei Ding, Rachel Chae, Jing Wang, Li Zhang, Hong Hong, Xiaohua Zhu, and Changzhi Li. Inattentive driving behavior detection based on portable fmcw radar. *IEEE Transactions on Microwave Theory and Techniques*, 67(10):4031–4041, 2019. *12, 14, 41, 115, 117, 121, 124, 125, 126, 127, 134, 138*

[67] Sidney K. D'Mello, Caitlin Mills, Robert Bixler, and Nigel Bosch. Zone out no more: Mitigating mind wandering during computerized reading. In *EDM*, 2017. *33, 57*

[68] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015. *9, 49*

[69] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. *51*

[70] Stephan Dreiseitl and Lucila Ohno-Machado. Logistic regression and artificial neural network classification models: a methodology review. *Journal of Biomedical Informatics*, 35(5):352–359, 2002. *43, 44*

[71] Hao Du, Yuan He, and Tian Jin. Transfer learning for human activities classification using micro-doppler spectrograms. In *2018 IEEE International Conference on Computational Electromagnetics (ICCEM)*, 03 2018. *115*

[72] Na Du, Feng Zhou, Elizabeth M. Pulver, Dawn M. Tilbury, Lionel P. Robert, Anuj K. Pradhan, and X. Jessie Yang. Predicting driver takeover performance in conditionally automated driving. *Accident Analysis & Prevention*, 148:105748, 2020. *102*

[73] Kaibo Duan, S Sathiya Keerthi, and Aun Neow Poo. Evaluation of simple performance measures for tuning svm hyperparameters. *Neurocomputing*, 51:41–59, 2003. *44*

[74] Andrew Duchowski. *Eye Tracking Methodology*. Springer, 05 2017. *37, 78, 79*

[75] Electron. *https://www.electronjs.org/*. *61*

[76] Baris Erol, Moeness Amin, Boualem Boashash, Fauzia Ahmad, and Yimin Zhang. Wideband radar based fall motion detection for a generic elderly. In *2016 50th Asilomar Conference on Signals, Systems and Computers*, pages 1768–1772. IEEE, 11 2016. *41*

[77] Gary W. Evans, Mary Anne Skorpanich, Tommy Gärling, Kendall J. Bryant, and Brian Bresolin. The effects of pathway configuration, landmarks and stress on environmental cognition. *Journal of Environmental Psychology*, 4(4):323–335, 1984. *24, 34, 35, 82, 84*

[78] Myrthe Faber and Sidney K. D'Mello. How the stimulus influences mind wandering in semantically rich task contexts. *Cognitive research: principles and implications*, 3(1):35, 2018. *32, 57, 58*

[79] Martha J. Farah. The neurological basis of mental imagery: A componential analysis. *Cognition*, 18(1):245–272, 1984. *21*

[80] Hartmut Feld, Bruno Mirbach, Jigyasa Singh Katrolia, Mohamed Selim, Oliver Wasenmüller, and Didier Stricker. Dfki cabin simulator: A test platform for visual in-cabin monitoring functions. In *Commercial Vehicle Technology 2020 - Proceedings of the 6th Commercial Vehicle Technology Symposium - CVT 2020. Commercial Vehicle Technology Symposium (CVT), 6th International Commercial Vehicle Technology Symposium Kaiserslautern, Kaiserlautern, Germany*, University of Kaiserslautern, 2020. University of Kaiserslautern, Springer. *87*

[81] Shi Feng and Gavin M. Bidelman. Music familiarity modulates mind wandering during lexical processing. In *CogSci*, 2015. *59*

[82] Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. Do we need hundreds of classifiers to solve real world classification problems? *The journal of machine learning research*, 15(1):3133–3181, 2014. *46*

[83] Francesco Fioranelli, Syed Aziz Shah, Haobo Li1, Aman Shrestha, Shufan Yang, and Julien Le Kernec. Radar sensing for healthcare. *Electronics Letters*, 55(19):1022–1024, 2019. *114*

[84] Robert J. Fontana. Recent system applications of short-pulse ultra-wideband (uwb) technology. *IEEE Transactions on Microwave Theory and Techniques*, 52(9):2087–2104, 2004. *41*

[85] Sophie Forster. Distraction and mind-wandering under load. *Frontiers in psychology*, 4:283, 05 2013. *58*

[86] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001. *46*

[87] Aurélien Géron. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. " O'Reilly Media, Inc.", 2022. *50, 51*

[88] Leonard Giambra and Alicia Grodsky. *Task-Unrelated Images and Thoughts While Reading*, pages 27–31. 01 1989. *58, 66*

[89] Christian Gold, D. Dambock, Lutz Lorenz, and Klaus Bengler. 'Take over!' How long does it take to get the driver back into the loop? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 57:1938–1942, 09 2013. *114*

[90] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016. *47*

[91] Klaus Gramann, Paul Hoepner, and Katja Karrer. Modified navigation instructions for spatial navigation assistance systems lead to incidental spatial learning. *Frontiers in Psychology*, 8, 02 2017. *35*

[92] Romain Grandchamp, Claire Braboszcz, and Arnaud Delorme. Oculometric variations during mind wandering. *Frontiers in Psychology*, 5:31, 2014. *32, 59, 78*

[93] Alberto Greco, Gaetano Valenza, Antonio Lanatà, Enzo Scilingo, and Luca Citi. cvxeda: A convex optimization approach to electrodermal activity processing. *IEEE Transactions on Biomedical Engineering*, 2016:797–804, 04 2016. *39, 40, 73, 74, 95*

[94] P. Green, W. Levison, G. Paelke, and C. Serafin. Preliminary human factors design guidlines for driver information systems. final report. Technical report, 1995. *84*

[95] Qiong Gu, Li Zhu, and Zhihua Cai. Evaluation measures of the classification performance of imbalanced data sets. In *International symposium on intelligence computation and applications*, pages 461–471. Springer, 2009. *53, 105*

[96] Tommy Gärling, Anders Böök, Erik Lindberg, and Tomas Nilsson. Memory for the spatial layout of the everyday physical environment: Factors affecting rate of acquisition. *Journal of Environmental Psychology*, 1(4):263–277, 1981. *35, 83*

[97] S. Hart. Nasa-task load index (nasa-tlx); 20 years later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 50:904 – 908, 2006. *84*

[98] Sandra G. Hart and Lowell E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In Peter A. Hancock and Najmedin Meshkati, editors, *Human Mental Workload*, volume 52 of *Advances in Psychology*, pages 139–183. North-Holland, 1988. *84*

[99] Alan M. Hay. The derivation of global estimates from a confusion matrix. *International Journal of Remote Sensing*, 9(8):1395–1398, 1988. *53*

[100] Francisco Hernando Gallego, David Luengo, and Antonio Artés Rodríguez. Feature extraction of galvanic skin responses by non-negative sparse deconvolution. *IEEE Journal of Biomedical and Health Informatics*, PP:1–1, 12 2017. *40*

[101] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. *49*

[102] Liberty Hoekstra-Atwood, David Prendez, John Campbell, and Christian Richard. Some on-road glances are more equal than others: Measuring engagement in the driving task. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 63:1986–1990, 11 2019. *103*

[103] Stephen Hutt, Jessica Hardey, Robert Bixler, Angela Stewart, Evan F. Risko, and Sidney D'Mello. Gaze-based detection of mind wandering during lecture viewing. In *EDM*, 2017. *33, 37*

[104] Stephen Hutt, Kristina Krasich, Caitlin Mills, Nigel Bosch, Shelby White, James Brockmole, and Sidney D'Mello. Automated gaze-based mind wandering detection during computerized learning in classrooms. *User Modeling and User-Adapted Interaction*, 29, 06 2019. *60*

[105] Stephen Hutt, Caitlin Mills, Shelby White, Patrick J Donnelly, and Sidney K D'Mello. The eyes have it: Gaze-based detection of mind wandering during learning with an intelligent tutoring system. ERIC, 2016. *33, 57*

[106] Unai Alegre Ibarra, Juan Carlos Augusto, and Carl Evans. Perspectives on engineering more usable context-aware systems. *Journal of Ambient Intelligence and Humanized Computing*, 9:1593–1609, 2018. *19, 24*

[107] InnoSenT. Incabin radar monitoring. *https://www.innosent.de/en/automotive/incabin-radar-monitoring/*. *41*

[108] Toru Ishikawa and Daniel Montello. Spatial knowledge acquisition from direct experience in the environment: Individual differences in the development of metric knowledge and the integration of separately learned places. *Cognitive psychology*, 52:93–129, 04 2006. *34, 35*

[109] Shoya Ishimaru, Syed Bukhari, Carina Heisel, Nicolas Großmann, Pascal Klein, Jochen Kuhn, and Andreas Dengel. *Augmented Learning on Anticipating Textbooks with Eye Tracking*, pages 387–398. Springer, 01 2018. *9, 20*

[110] P. G. Jackson. How will route guidance information affect cognitive maps? *Journal of Navigation*, 49(2):178–186, 1996. *21, 35*

[111] Chi Jian-Nan, Zhang Peng-Yi, Zheng Si-Yi, Zhang Chuang, and Huang Ying. Key techniques of eye gaze tracking based on pupil corneal reflection. In *2009 WRI Global Congress on Intelligent Systems*, volume 2, pages 133–138, 2009. *37*

[112] Helen Jing, Karl Szpunar, and Daniel Schacter. Interpolated testing influences focused attention and improves integration of information during a video-recorded lecture. *Journal of experimental psychology. Applied*, 22, 06 2016. *32*

[113] jMonkeyEngine. *https://jmonkeyengine.org/*. *89*

[114] Ida Arlene Joiner. Chapter 4 – driverless vehicles: Pick me up at the...? In Ida Arlene Joiner, editor, *Emerging Library Technologies*, Chandos Information Professional Series, pages 69–94. Chandos Publishing, 2018. *114*

[115] Branka Jokanovic, Moeness Amin, and Fauzia Ahmad. Radar fall motion detection using deep learning. In *2016 IEEE radar conference (RadarConf)*, pages 1–6, 05 2016. *41*

[116] Julia Kam. Slow fluctuations in attentional control of sensory cortex. *Journal of Cognitive Neuroscience*, 23:460–470, 01 2011. *32*

[117] Michael Kane, Bridget Smeekens, Claudia von Bastian, John Lurquin, Nicholas Carruth, and Akira Miyake. A combined experimental and individual-differences investigation into mind wandering during a video lecture. *Journal of Experimental Psychology General*, in press, 07 2017. *32, 57*

[118] Bronisław Kapitaniak, Marta Walczak, Marcin Kosobudzki, Zbigniew Jóźwiak, and Alicja Bortkiewicz. Application of eye-tracking in drivers testing: A review of research. *International Journal of Occupational Medicine and Environmental Health*, 28(6):941–954, 2015. *20, 24, 34, 37, 82*

[119] Jakob Karolus and Paweł W Woźniak. Proficiency-aware systems: Designing for user reflection in context-aware systems. *it-Information Technology*, 63(3):167–175, 2021. *19*

[120] Jigyasa Katrolia, Bruno Mirbach, Ahmed El-Sherif, Hartmut Feld, Jason Rambach, and Didier Stricker. Ticam: A time-of-flight in-car cabin monitoring

dataset. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2021. *114*

[121] Rebecca Keogh and Joel Pearson. The blind mind: No sensory visual imagery in aphantasia. *Cortex*, 105:53–60, 2018. The Eye's Mind - visual imagination, neuroscience and the humanities. *21*

[122] Khan and Lee. Gaze and eye tracking: Techniques and applications in adas. *Sensors*, 19:5540, 12 2019. *41*

[123] Beob Kim and Hans Stein. A spreadsheet program for making a balanced latin square design. *Revista Colombiana de Ciencias Pecuarias*, 22:591–596, 10 2009. *92*

[124] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Neural Information Processing Systems (NIPS)*, 2012. *48*

[125] Benjamin Kuipers. Modeling spatial knowledge. *Cognitive Science*, 2(2):129–153, 1978. *34*

[126] Moritz Körber, Andrea Cingel, Markus Zimmermann, and Klaus Bengler. Vigilance decrement and passive fatigue caused by monotony in automated driving. *Procedia Manufacturing*, 3:2403–2409, 12 2015. *34*

[127] Richard J. Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33 1:159–74, 1977. *60*

[128] Frederic Lardinois. BMW launches gaze detection so your car knows what you're looking at. *https://techcrunch.com/2020/01/07/bmw-launches-gaze-detection-so-your-car-knows-what-youre-looking-at/*, January 2020. *24, 82, 103*

[129] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. *9, 43, 47, 48*

[130] Seong Leem, Faheem Khan, and Sung Ho Cho. Vital sign monitoring and mobile phone usage detection using ir-uwb radar for intended use in car crash prevention. *Sensors*, 17:1240, 05 2017. *41, 114*

[131] Jing Li, Zhaofa Zeng, Jiguang Sun, and Fengshan Liu. Through-wall detection of human being's movement by uwb radar. *IEEE Geoscience and Remote Sensing Letters*, 9(6):1079–1083, 2012. *117*

[132] Xinyu Li, Yuan He, and Xiaojun Jing. A survey of deep learning-based human activity recognition in radar. *Remote Sensing*, 11(9), 2019. *41, 42*

[133] Xinyu Li, Yuan He, Yang Yang, Yuanquan Hong, and Xiaojun Jing. LSTM based human activity classification on radar range profile. In *2019 IEEE International Conference on Computational Electromagnetics (ICCEM)*, pages 1–2. IEEE, 2019. *116*

[134] Nade Liang, Jing Yang, Denny Yu, Kwaku Prakah-Asante, Reates Curry, Mike Blommer, Radhakrishnan Swaminathan, and Brandon Pitts. Using eye-tracking to investigate the effects of pre-takeover visual engagement on situation awareness during automated driving. *Accident Analysis & Prevention*, 157:106143, 07 2021. *84, 103, 114*

[135] Yulan Liang, John Lee, and Lora Yekhshatyan. How dangerous is looking away from the road? algorithms predict crash risk from glance patterns in naturalistic driving. *Human factors*, 54:1104–16, 12 2012. *83, 121*

[136] Lisa K. Libby and Richard P. Eibach. Chapter four - visual perspective in mental imagery: A representational tool that functions in judgment, emotion, and self-insight. volume 44 of *Advances in Experimental Social Psychology*, pages 185–245. Academic Press, 2011. *21*

[137] Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018. *54*

[138] Monika Lohani, Brennan R. Payne, and David L. Strayer. A review of psychophysiological measures to assess cognitive states in real-world driving. *Frontiers in Human Neuroscience*, 13:57, 2019. *84*

[139] Tyron Louw and Natasha Merat. Are you in the loop? Using gaze dispersion to understand driver visual attention during vehicle automation. *Transportation Research Part C Emerging Technologies*, 76:35–50, 03 2017. *85, 114*

[140] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pages 4765–4774, 2017. *10, 54, 77, 108*

[141] Heinrich Löwen, Jakub Krukar, and Angela Schwering. Spatial learning with orientation maps: The influence of different environmental features on spatial knowledge acquisition. *International Journal of Geo-Information*, 8:149, 03 2019. *34, 35*

[142] Wilson E. Marcílio and Danilo M. Eler. From explanations to feature selection: assessing shap values as feature selection mechanism. In *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 340–347, 2020. *54*

[143] Manuel Martin, Alina Roitberg, Monica Haurilet, Matthias Horne, Simon Reiß, Michael Voit, and Rainer Stiefelhagen. Drive&act: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2801–2810, 2019. *114*

[144] Raubal Martin and Stephan Winter. Enriching wayfinding instructions with local landmarks. *Geographic Information Science*, pages 243–259, 01 2002. *92*

[145] Marco Mercuri, Yao-Hong Liu, Ilde Lorato, Tom Torfs, Fokko Wieringa, André Bourdoux, and Chris Van Hoof. A direct phase-tracking doppler radar using wavelet independent component analysis for non-contact respiratory and heart rate monitoring. *IEEE Transactions on Biomedical Circuits and Systems*, 12(3):632–643, 2018. *41*

[146] Negin Minaei. Do modes of transportation and gps affect cognitive maps of londoners? *Transportation research part A: policy and practice*, 70:162–180, 2014. *83*

[147] Hyundai MOBIS. The new radar-based occupant alert system to keep your children safe. *https://www.hyundaimotorgroup.com/story/CONT0000000000002988*. *41*

[148] Andrew Mondschein, Evelyn Blumenberg, and Brian Taylor. Accessibility and cognition: The effect of transportation mode on spatial knowledge. *University of California Transportation Center, University of California Transportation Center, Working Papers*, 01 2008. *34, 35, 83*

[149] Walter Morales-Alvarez, Oscar Sipele, Régis Léberon, Hadj Tadjine, and Cristina Olaverri Monreal. Automated driving: A literature review of the take over request in conditional automation. *Electronics*, 9:2087, 12 2020. *24, 82, 83, 103*

[150] Drew Morris, Jason Erno, and June Pilcher. Electrodermal response and automation trust during simulated self-driving car use. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 61:1759–1762, 09 2017. *35, 84*

[151] Stefan Münzer, Hubert Zimmer, Maximilian Schwalm, Jörg Baus, and Ilhan Aslan. Computer-assisted navigation and the acquisition of route and survey knowledge. *Journal of Environmental Psychology*, 26:300–308, 12 2006. *35*

[152] Alexey Natekin and Alois Knoll. Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, 7:21, 12 2013. *46, 47*

[153] Farzan M Noori, Md Zia Uddin, and Jim Torresen. Ultra-wideband radar-based activity recognition using deep learning. *IEEE Access*, 9:138132–138143, 2021. *41, 115, 117, 123*

[154] Novelda. Legacy-sw. *https://github.com/novelda/Legacy-SW*, 2022. *118*

[155] On-Road Automated Driving (ORAD) committee. *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*, jun 2018. *9, 33*

[156] OpenDs. *https://opends.dfki.de*, 2022. *89, 119*

[157] Mahesh Pal. Random forest classifier for remote sensing classification. *International journal of remote sensing*, 26(1):217–222, 2005. *44*

[158] Raja Parasuraman and Dietrich Manzey. Complacency and bias in human use of automation: An attentional integration. *Human factors*, 52:381–410, 06 2010. *84*

[159] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. Specaugment: A simple data augmentation method for automatic speech recognition. *Interspeech 2019*, Sep 2019. *134*

[160] Avi Parush, Shir Ahuvia, and Ido Erev. Degradation in spatial knowledge acquisition when using automatic navigation systems. In *COSIT*, 2007. *84*

[161] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 12 2019. *122*

[162] Joel Pearson. The human imagination: the cognitive neuroscience of visual mental imagery. *Nature reviews neuroscience*, 20(10):624–634, 2019. *21*

[163] Joel Pearson, Rosanne L Rademaker, and Frank Tong. Evaluating the mind's eye: the metacognition of visual imagery. *Psychological Science*, 22(12):1535–1542, 2011. *22*

[164] Matthew H. Phillips and Jay A. Edelman. The dependence of visual scanning performance on search direction and difficulty. *Vision Research*, 48(21):2184–2192, 2008. *112*

[165] Benjamin Plimpton, Priya Patel, and Lia Kvavilashvili. Role of triggers and dysphoria in mind-wandering about past, present and future: A laboratory study. *Consciousness and cognition*, 33C:261–276, 02 2015. *22*

[166] Alan T Pope, Edward H Bogart, and Debbie S Bartolome. Biocybernetic system evaluates indices of operator engagement in automated task. *Biological Psychology*, 40(1):187–195, 1995. EEG in Basic and Applied Settings. *34*

[167] Psychometrica. *https://www.psychometrica.de/lix.html*. *64*

[168] Yue Qin and Hassan A. Karimi. Spatial knowledge acquisition for cognitive maps in autonomous vehicles. In Don Harris and Wen-Chin Li, editors, *Engineering Psychology and Cognitive Ergonomics. Cognition and Design*, pages 384–397, Cham, 2020. Springer International Publishing. *35, 81, 83*

[169] Chongxiao Qu, Yongjin Zhang, Lei Jin, Changjun Fan, Shuo Liu, and Xiayan Chen. Exploring hand gesture recognition using micro-doppler radar data based on vision transformers. *Journal of Physics: Conference Series*, 2504(1):012046, may 2023. *116*

[170] Marcus E. Raichle, Ann Mary MacLeod, Abraham Z. Snyder, William J. Powers, Debra A. Gusnard, and Gordon L. Shulman. A default mode of brain function. *Proceedings of the National Academy of Sciences*, 98(2):676–682, 2001. *32*

[171] Kishore Ramaiah. In-cabin radar can sense children in second- and third-row vehicles. *https://www.electronicproducts.com/in-cabin-radar-can-sense-children-in-second-and-third-row-vehicles/*. *41*

[172] Keith Rayner, Timothy J Slattery, and Nathalie N Bélanger. Eye movements, the perceptual span, and reading speed. *Psychonomic bulletin & review*, 17(6):834–839, 2010. *38*

[173] Michael A. Regan and Charlene Hallett. Chapter 20 - driver distraction: Definition, mechanisms, effects, and mitigation. In Bryan E. Porter, editor, *Handbook of Traffic Psychology*, pages 275–286. Academic Press, San Diego, 2011. *120*

[174] Bryan Reimer, Anthony Pettinato, Bobbie Seppelt, Lex Fridman, Joonbum Lee, Junghee Park, Karl Iagnemma, and Bruce Mehler. Behavioral impact of drivers' roles in automated driving. In *Proceedings of the 8th international conference on automotive user interfaces and interactive vehicular applications*, pages 217–224, 10 2016. *34*

[175] Fritz Renner, Fionnuala C. Murphy, Julie L. Ji, Tom Manly, and Emily A. Holmes. Mental imagery as a "motivational amplifier" to promote activities. *Behaviour Research and Therapy*, 114:51–59, 2019. *36*

[176] Alexander T Sack, Julia M Sperling, David Prvulovic, Elia Formisano, Rainer Goebel, Francesco Di Salle, Thomas Dierks, and David EJ Linden. Tracking the mind's image in the brain ii: transcranial magnetic stimulation reveals parietal asymmetry in visuospatial imagery. *Neuron*, 35(1):195–204, 2002. *20, 21, 23, 34*

[177] Umer Saeed, Syed Yaseen Shah, Abdullah Alhumaidi Alotaibi, Turke Althobaiti, Naeem Ramzan, Qammer H. Abbasi, and Syed Aziz Shah. Portable UWB RADAR sensing system for transforming subtle chest movement into actionable micro-doppler signatures to extract respiratory rate exploiting resnet algorithm. *IEEE Sensors*, 21(20):23518–23526, 2021. *9, 41, 118*

[178] Takuya Sakamoto. Personal identification using ultrawideband radar measurement of walking and sitting motions and a convolutional neural network. *arXiv preprint arXiv:2008.02182*, 08 2020. *117*

[179] Diego Alejandro Salazar, Jorge Iván Vélez, and Juan Carlos Salazar. Comparison between svm and logistic regression: Which one is better to discriminate? *Revista Colombiana de Estadística*, 35(SPE2):223–237, 2012. *44, 45, 76*

[180] Dario Salvucci and Joseph Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the Eye Tracking Research and Applications Symposium*, pages 71–78, 01 2000. *72*

[181] Jonathan Schooler, Jonathan Smallwood, Kalina Christoff, Todd Handy, Erik Reichle, and Michael Sayette. Meta-awareness, perceptual decoupling and the wandering mind. *Trends in cognitive sciences*, 15:319–26, 06 2011. *32, 59*

[182] Paul Seli, Evan Risko, Daniel Smilek, and Daniel Schacter. Mind-wandering with and without intention. *Trends in Cognitive Sciences*, 20, 06 2016. *22, 32*

[183] Yuming Shao, Sai Guo, Lin Sun, and Weidong Chen. Human motion classification based on range information with deep convolutional neural network. In *2017*

4th International Conference on Information Science and Control Engineering (ICISCE), pages 1519–1523, 07 2017. *41, 115, 117, 123*

[184] Lloyd S Shapley. A value for n-person games. *Classics in game theory*, 69, 1997. *54*

[185] Roger N. Shepard and Jacqueline Metzler. Mental rotation of three-dimensional objects. *Science*, 171(3972):701–703, 1971. *32*

[186] Yuan Shi, Jeyhoon Maskani, Giandomenico Caruso, and Monica Bordegoni. Explore user behaviour in semi-autonomous driving. *Proceedings of the Design Society: International Conference on Engineering Design*, 1:3871–3880, 07 2019. *34*

[187] Aman Shrestha, Haobo Li, Julien le kernec, and Francesco Fioranelli. Continuous human activity classification from fmcw radar with bi-lstm networks. *IEEE Sensors Journal*, PP:1–1, 07 2020. *117*

[188] Alexander W. Siegel and Sheldon H. White. The development of spatial representations of large-scale environments. *Advances in child development and behavior*, 10:9–55, 1975. *34*

[189] Sketchfab. *https://sketchfab.com/3d-models*. *91*

[190] Jonathan Smallwood, Kevin S. Brown, Christine Tipper, Barry Giesbrecht, Michael S. Franklin, Michael D. Mrazek, Jean M. Carlson, and Jonathan W. Schooler. Pupillometric evidence for the decoupling of attention from perceptual input during offline thought. *PLOS ONE*, 6(3):1–8, 03 2011. *32, 37, 59*

[191] Jonathan Smallwood, John B Davies, Derek Heim, Frances Finnigan, Megan Sudberry, Rory O'Connor, and Marc Obonsawin. Subjective experience and the attentional lapse: Task engagement and disengagement during sustained attention. *Consciousness and cognition*, 13(4):657–690, 2004. *32*

[192] Jonathan Smallwood, Daniel J. Fishman, and Jonathan W. Schooler. Counting the cost of an absent mind: Mind wandering as an underrecognized influence on

educational performance. *Psychonomic bulletin & review*, 14(2):230–236, 2007. *31, 32*

[193] Jonathan Smallwood, Louise Nind, and Rory C. O'Connor. When is your head at? an exploration of the factors associated with the temporal focus of the wandering mind. *Consciousness and cognition*, 18(1):118–125, 2009. *58*

[194] Jonathan Smallwood and Jonathan Schooler. The restless mind. *Psychology of Consciousness: Theory, Research, and Practice*, 1:130–149, 08 2013. *32, 58*

[195] Jonathan Smallwood and Jonathan W. Schooler. The science of mind wandering: Empirically navigating the stream of consciousness. *Annual Review of Psychology*, 66(1):487–518, 2015. *21, 31, 58*

[196] Daniel Smilek, Jonathan S. A. Carriere, and J. Allan Cheyne. Out of mind, out of sight: eye blinking as indicator and embodiment of mind wandering. *Psychological science*, 21(6):786–789, 2010. *32, 59*

[197] Philip Smith and Daniel Little. Small is beautiful: In defense of the small-n design. *Psychonomic Bulletin & Review*, 25, 03 2018. *86, 128*

[198] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45:427–437, 07 2009. *105*

[199] Sandro Sperandei. Understanding logistic regression analysis. *Biochemia medica*, 24:12–8, 02 2014. *43*

[200] Alexander Statnikov and Constantin F Aliferis. Are random forests better than support vector machines for microarray-based cancer classification? In *AMIA annual symposium proceedings*, volume 2007, page 686. American Medical Informatics Association, 2007. *76*

[201] David Stawarczyk, Steve Majerus, Pierre Maquet, and Arnaud D'Argembeau. Neural correlates of ongoing conscious experience: Both task-unrelatedness and stimulus-independence are related to default network activity. In *PloS one*, 2011. *32*

[202] Lena Steindorf and Jan Rummel. Do your eyes give you away? a validation study of eye-movement measures used as indicators for mindless reading. *Behavior Research Methods*, 52, 02 2019. *21, 31, 32, 59*

[203] Jill C. Stoltzfus. Logistic regression: A brief primer. *Academic Emergency Medicine*, 18(10):1099–1104, 2011.

[204] Liila Taruffi, Corinna Pehrs, Stavros Skouras, and Stefan Koelsch. Effects of sad and happy music on mind-wandering and the default mode network. *Scientific Reports*, 7(1):14396, 2017. *32, 58, 59, 61, 64*

[205] William Taylor, Kia Dashtipour, Syed Aziz Shah, Amir Hussain, Qammer H. Abbasi, and Muhammad A. Imran. Radar sensing for activity classification in elderly people exploiting micro-doppler signatures using machine learning. *Sensors*, 21(11), 2021. *41, 115, 117, 123, 134, 136*

[206] Brad Templeton. New tesla autopilot statistics show it's almost as safe driving with it as without. *https://www.forbes.com/sites/bradtempleton/2020/10/28/new-tesla-autopilot-statistics-show-its-almost-as-safe-driving-with-it-as-without/*, October 2020. *82, 114*

[207] Tesla. Car safety and security features. *https://www.tesla.com/support/car-safety-security-features*. *82*

[208] Nigel JT Thomas. Are theories of imagery theories of imagination? an active perception approach to conscious mental content. *Cognitive science*, 23(2):207–245, 1999. *20*

[209] Perry W Thorndyke and Barbara Hayes-Roth. Differences in spatial knowledge acquired from maps and navigation. *Cognitive Psychology*, 14(4):560–589, 1982. *81, 103*

[210] Florentin Thullier, Julien Maitre, Sebastien Gaboury, and Kévin Bouchard. A systematic evaluation of the xethru x4 ultra-wideband radar behavior. *Procedia Computer Science*, 198:148–155, 01 2022. *118*

[211] Emma Tivesten and Marco Dozza. Driving context and visual-manual phone tasks influence glance behavior in naturalistic driving. *Transportation Research Part F: Traffic Psychology and Behaviour*, 26:258–272, 2014. *37*

[212] Edward C Tolman. Cognitive maps in rats and men. *Psychological review*, 55(4):189, 1948. *34*

[213] Carnegie Mellon University. CMU Graphics Lab Motion Capture Database. *http://mocap.cs.cmu.edu/*. *116*

[214] Sarah Uzzaman and Steve Joordens. The eyes know what you are thinking: Eye movements as an objective measure of mind wandering. *Consciousness and cognition*, 20(4):1882–1886, 2011. *59, 78*

[215] Karl F Van Orden, Wendy Limbert, Scott Makeig, and Tzyy-Ping Jung. Eye activity correlates of workload during a visuospatial memory task. *Human factors*, 43(1):111–121, 2001. *38*

[216] Dana van Son, Frances De Blasio, Jack Fogarty, Angelos Angelidis, Robert Barry, and Peter Putman. Frontal eeg theta/beta ratio during mind wandering episodes. *Biological Psychology*, 140, 11 2018. *21, 31, 32*

[217] Hendrik A. H. C. van Veen, Hartwig K. Distler, Stephan J. Braun, and Heinrich H. Bülthoff. Navigating through a virtual city: Using virtual reality technology to study human action and perception. *Future Gener. Comput. Syst.*, 14:231–242, 1998. *34*

[218] Jake VanderPlas. *Python data science handbook: Essential tools for working with data*. " O'Reilly Media, Inc.", 2016. *45*

[219] Baptist Vandersmissen, Nicolas Knudde, Azarakhsh Jalalvand, Ivo Couckuyt, Tom Dhaene, and Wesley De Neve. Indoor human activity recognition using high-dimensional sensors and deep neural networks. *Neural Computing and Applications*, 32, 08 2020. *41, 117*

[220] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need.

In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. *9, 50, 51*

[221] Trent W. Victor, Emma Tivesten, Pär Gustavsson, Joel Johansson, Fredrik Sangberg, and Mikael Ljung Aust. Automation expectation mismatch: Incorrect prediction despite eyes on threat and hands on wheel. *Human Factors*, 60(8):1095–1116, 2018. PMID: 30096002. *82, 83, 103, 114, 121*

[222] Volvo Cars. Volvo cars to deploy in-car cameras and intervention against intoxication, distraction. *https://www.media.volvocars.com/global/en-gb/media/pressreleases/250015/volvo-cars-to-deploy-in-car-cameras-and-intervention-against-intoxication-distraction*, March 2019. *82, 114*

[223] Rul von Stülpnagel and Melanie Steffens. Can active navigation be as good as driving? a comparison of spatial memory in drivers and backseat drivers. *Journal of experimental psychology. Applied*, 18:162–77, 02 2012. *35, 83, 84*

[224] Sonja Walcher, Christof Korner, and Mathias Benedek. Looking for ideas: Eye behavior during goal-directed internally focused cognition. *Consciousness and cognition*, 53:165–175, 2017. *32, 59*

[225] Chengshun Wang, Yufen Chen, Shulei Zheng, Yecheng Yuan, and Wang Shuang. Research on generating an indoor landmark salience model for self-location and spatial orientation from eye-tracking data. *ISPRS International Journal of Geo-Information*, 9:97, 02 2020. *34*

[226] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 568–578, 2021. *129*

[227] Zhaoyue Wang, Sha Huan, Limei Wu, Qingyuan Wang, Jiajun Liu, and Zegui Hu. Attention-based vision transformer for human activity classification using mmwave radar. In *Proceedings of the 2022 4th International Conference on*

Video, Signal and Image Processing, VSIP '22, page 128–134, New York, NY, USA, 2023. Association for Computing Machinery. *51, 116, 117*

[228] Yana Weinstein. Mind-wandering, how do i measure thee with probes? let me count the ways. *Behavior research methods*, 50(2):642–661, 2018. *32, 66*

[229] Wisnu Wiradhany, Marieke Vugt, and Mark Nieuwenstein. Media multitasking, mind-wandering, and distractibility: A large-scale study. *Attention, Perception, & Psychophysics*, 08 2019. *58*

[230] Wissen.de. *https://www.wissen.de/*. *63*

[231] Shiyan Yang, Jonny Kuo, and Michael Lenné. Effects of distraction in on-road level 2 automated driving: Impacts on glance behavior and takeover performance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, page 001872082093679, 07 2020. *41, 114, 117, 123*

[232] Yang Yang, Chunping Hou, Yue Lang, Dai Guan, Danyang Huang, and Jinchen Xu. Open-set human activity recognition based on micro-doppler signatures. *Pattern Recognition*, 85:60–69, 2019. *41*

[233] Alfred L Yarbus. *Eye movements and vision*. Springer, 2013. *37*

[234] Alexander G. Yarovoy, Leo P. Ligthart, Jonas. Matuzas, and Boris Levitas. UWB radar for human being detection. *IEEE Aerospace and Electronic Systems Magazine*, 21(3):10–14, 2006. *41*

[235] Johannes Zagermann, Ulrike Pfeil, and Harald Reiterer. Measuring cognitive load using eye tracking technology in visual computing. In *Proceedings of the sixth workshop on beyond time and errors on novel evaluation methods for visualization*, pages 78–85, 2016. *37*

[236] Maryam Zahabi, Yinsong Wang, and Shahin Shahrampour. Classification of officers' driving situations based on eye-tracking and driver performance measures. *IEEE Transactions on Human-Machine Systems*, 51(4):394–402, 2021. *102, 103, 111*

[237] Kathrin Zeeb, Axel Buchner, and Michael Schrauf. Is take-over time all that matters? The impact of visual-cognitive load on driver take-over quality after conditionally automated driving. *Accident; analysis and prevention*, 92:230–239, 04 2016. *114*

[238] Raimondas Zemblys, Diederick Niehorster, Oleg Komogortsev, and Kenneth Holmqvist. Using machine learning to detect events in eye-tracking data. *Behavior Research Methods*, 50, 02 2017. *101, 104*

[239] Zehui Zhan, Lei Zhang, Hu Mei, and Patrick SW Fong. Online learners' reading ability detection based on eye-tracking sensors. *Sensors*, 16(9):1457, 2016. *38*

[240] Cemin Zhang, Michael Kuhn, B. Merkl, Aly Fathy, and Mohamed Mahfouz. Accurate uwb indoor localization system utilizing time difference of arrival approach. In *2006 IEEE radio and wireless symposium*, pages 515 – 518, 02 2006. *42*

[241] Shangyue Zhu, Junhong Xu, Hanqing Guo, Qiwei Liu, Shaoen Wu, and Honggang Wang. Indoor human activity recognition based on ambient radar with signal processing and machine learning. In *2018 IEEE International Conference on Communications (ICC)*, pages 1–6, 2018. *41*

Curriculum Vitae

Iuliia Brishtel

Education

RPTU Kaiserslautern - Landau, Kaiserslautern, Germany
Doctor of Engineering, Computer Science 2019 - 2024
TU Kaiserslautern, Kaiserslautern, Germany
Master of Science, Cognitive Science 2016 - 2018
TU Kaiserslautern, Kaiserslautern, Germany
Bachelor of Arts, Applied Social Sciences 2013 - 2016

Professional Experience

Boehringer Ingelheim Pharma GmbH & Co. KG, Ingelheim am Rhein, Germany
Postdoctoral Researcher, Biostatistics and Data Science 2023 - Present
Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach an der Riss, Germany
Intern, Biostatistics and Data Science 2022 - 2023
German Research Center for Artificial Intelligence, Kaiserslautern, Germany
Researcher, Augmented Vision 2018 - 2023
German Research Center for Artificial Intelligence, Kaiserslautern, Germany
Working Student, Smart Data & Knowledge Services 2017 - 2018