Moritz Langner*, Peyman Toreini and Alexander Maedche

# Cognitive state detection with eye tracking in the field: an experience sampling study and its lessons learned

https://doi.org/10.1515/icom-2023-0035
Received November 24, 2023; accepted March 21, 2024;
published online April 15, 2024

**Abstract:** In the future, cognitive activity will be tracked in the same way how physical activity is tracked today. Eye-tracking technology is a promising off-body technology that provides access to relevant data for cognitive activity tracking. For building cognitive state models, continuous and longitudinal collection of eye-tracking and self-reported cognitive state label data is critical. In a field study with 11 students, we use experience sampling and our data collection system esmLoop to collect both cognitive state labels and eye-tracking data. We report descriptive results of the field study and develop supervised machine learning models for the detection of two eye-based cognitive states: cognitive load and flow. In addition, we articulate the lessons learned encountered during data collection and cognitive state model development to address the challenges of building generalizable and robust user models in the future. With this study, we contribute knowledge to bring eye-based cognitive state detection closer to real-world applications.

**Keywords:** experience sampling; cognitive states; eye tracking; cognitive load; flow; machine learning

## 1 Introduction

Similar to physical activities such as steps per day that can be easily detected by wearables nowadays, it should become possible to detect cognitive user states like cognitive load, flow or mind wandering in the future. Cognition in general refers to the mental processes related to the acquisition, organization, and use of knowledge covering attention, memory, reasoning, decision making and problem solving.[1,2] Recently, there has been a call for research aimed at automatically detecting cognitive user states from on-body and off-body technologies.[3,4] Following the paradigm of biosignal-adaptive systems described in Schultz and Maedche,[5] the detection of cognitive user states based on biosignal data would allow to design interactive systems that adapt to the user's current needs, ultimately improving the user's performance and well-being. For example, in learning, the ability to detect cognitive states such as cognitive load, mind wandering, flow or situation awareness can be used to personalize learning content for the learner.[6,7] In the workplace, this capability can help to design work environments that enhance employee performance and well-being (e.g., through flow-adaptive notification management, load-adaptive task assignment, adaptive video meeting systems[8] or visual attention feedback[9,10]). By recognizing the user's cognitive state, interactive systems can not only better adapt to their users in a specific situation, but may also support users to learn from their past cognitive states in relation to their behavior. For example, users could adjust their work schedules based on identified patterns in the individual cognitive demands of specific tasks.

In particular, commercial off-the-shelf (COTS) eye trackers are a promising off-body technology that can provide access to cognitive user states.[6] Advances in eye-tracking technology combined with supervised machine learning (ML) have demonstrated that it is possible to identify cognitive user states based on collected eye data.[6,11] Current studies investigating cognitive state detection using eye-tracking technology and supervised ML have focused on collecting data from well-defined tasks in laboratory environments.[12–14] However, the resulting cognitive state ML models have the drawback of requiring data collected in highly controlled environments. To move these cognitive state models out of the laboratory and into real-world applications, the models must not only be accurate, but also robust and generalizable across tasks, users, and environments. To achieve high generalizability and robustness, data and labels must be collected for a variety of tasks and from users in different environments, ideally continuously over time. So far, only few studies have investigated the

***Corresponding author: Moritz Langner**, Karlsruhe Institute of Technology, Karlsruhe, Germany, E-mail: moritz.langner@kit.edu. https://orcid.org/0000-0001-7860-7118
**Peyman Toreini and Alexander Maedche**, Karlsruhe Institute of Technology, Karlsruhe, Germany, E-mail: peyman.toreini@kit.edu (P. Toreini), alexander.maedche@kit.edu (A. Maedche). https://orcid.org/0000-0002-2468-1715 (P. Toreini), https://orcid.org/0000-0001-6546-4816 (A. Maedche)

development of eye-based cognitive state ML models outside the laboratory or used eye-tracking data recorded over a longer period of time.[6,15] Thus, this remains an important research gap.

The Experience Sampling Method (ESM) is an established method for building user models and collecting a variety of data at random times and locations.[16] In ESM studies, participants are interrupted randomly, event-based, or interval-based to complete a self-report survey, typically in the form of questionnaires.[17] Therefore, ESM is a common method for systematically capturing people's activities, emotions, and thoughts during their daily lives based on self-reported information.[18] In addition, data from sensors such as GPS, gyroscope, temperature, air pressure, or electrocardiography are continuously collected and later correlated with the self-reported survey data.[19] ESM can also be used in combination with eye tracking to collect self-reported cognitive state data, also called labels in this context, to develop more robust, accurate, and generalizable eye-based cognitive state models. Such labels, in combination with eye-tracking data, serve as input for supervised ML algorithms to predict the corresponding cognitive state of the user. Since eye trackers are sensitive to changes in the environment and eye movements are task dependent, guidelines on how to apply eye tracking in ESM studies and develop eye-based cognitive state models based on label data collected in the field are crucial. However, guidelines for applying eye tracking in ESM studies are lacking, and existing ESM knowledge should be extended to an eye tracking context.

In this paper, we investigate whether it is feasible to use an ESM approach to collect cognitive state labels and eye-tracking data in the field as a foundation for the development of accurate, generalizable, and robust cognitive state models. We conduct an exploratory longitudinal ESM study to collect eye-tracking data in the field and develop eye-based cognitive state models on this basis. First, we introduce our experience sampling-based system, esmLoop, which is designed to collect eye-tracking data continuously and cognitive state labels randomly. We then outline the various steps required in preparing and conducting our ESM field study, which involved collecting data from 11 students working on their thesis project over 5 days using esmLoop. We provide detailed insights into the participants' interaction with esmLoop during the field study and their opinions and needs regarding this system, as articulated in post-study interviews. We focus on two exemplary cognitive states for supervised ML model development, namely cognitive load and flow. First, we develop supervised ML models following a classification- and regression-based approach for both cognitive states, using all collected label and eye-tracking data from different window sizes. Since the models do not significantly outperform baseline models, this highlights the challenge of developing eye-based ML models that are generalizable across tasks and participants, even though we collect a reasonable amount of labels for different tasks. Subsequently, we focus solely on the labels obtained during writing tasks to investigate the feasibility of building task-focused cognitive load and flow models based on eye-tracking data collected in the field. These writing task-focused models ultimately outperform the baseline models, demonstrating the feasibility of developing accurate and robust cognitive state models that are generalizable across participants using eye-tracking data collected in the field. Finally, we present the lessons learned during ESM data collection and cognitive state model development to address our challenges in building generalizable and robust models in the future. By sharing our experiences and providing the aggregated data and analysis scripts according to the open science paradigm, we fill the introduced research gap of using ESM to collect eye tracking and cognitive state labels in the field and develop eye-based cognitive state models based on it, thus contributing to the field of eye-tracking research. We believe our experience will facilitate the integration of eye tracking and eye-based cognitive state detection into real-world applications. Additionally, it can aid researchers in designing better ESM studies and developing cognitive state models in the future.

# 2 Related work

## 2.1 Eye-based recognition of user states: cognitive load and flow

The saying "the eyes are the window to our soul" highlights that the eyes provide more information about us humans than just visual attention.[20] Eye tracking is a technology that can provide access to much more user information, in particular with regards to cognitive user states. Leveraging the eye-mind hypothesis by Just and Carpenter,[21] fixations are directly reflecting what humans are cognitively processing. In recent years, eye-tracking technology has advanced a lot, especially in terms of robustness, so that it has made its way out of research laboratories into real-world applications.[22] With advances in computing power and machine learning algorithms, new approaches to eye-tracking data analysis have become possible. Support vector machines, k-nearest-neighbor and random forests are the most commonly applied machine learning algorithms in

combination with eye-tracking data for user state and trait recognition.[23–26] In particular, fixation, saccade and pupil-based features are used in both low-level and high-level (AOI-based) gaze features.

A recent trend is the investigation and detection of user traits and characteristics as well as cognitive and affective user states using eye-tracking technology. Research focused on the detection of user characteristics such as personality, working memory and field dependence based on eye-tracking data.[24,27,28] Studies targeting the recognition of affective user states using eye tracking focus mainly on the arousal and valence dimensions of emotions or on discrete emotions of Ekman.[29,30] Typical cognitive user states studied using eye-tracking technology are cognitive load and mind wandering.[6,11,31]

In our study we specifically focus on two cognitive states: cognitive load and flow. While cognitive load is already researched with eye-tracking technology, flow to the best of our knowledge was not yet investigated using eye-tracking technology. Cognitive load refers to how many mental resources are currently occupied and is typically captured by the NASA TLX.[32] The NASA TLX covers cognitive load in terms of six dimensions: mental demand, physical demand, temporal demand, performance, effort and frustration. Kahneman and Beatty[33] established the link between pupil size and cognitive load and several studies further investigated this link.[34,35] Many studies that predicted cognitive load using machine learning rely heavily on pupil-based metrics such as pupil dilation or blinking.[11,14,25] However, pupil size is not only dependent on the person's current cognitive load, but also influenced by other environmental factors such as the ambient light conditions. Therefore, cognitive load recognition in the field requires more elaborated approaches than just measuring pupil size. Recent publications specifically investigated cognitive load recognition including further typical eye-tracking features such as fixations, saccades or microsaccades.[3,36,37] In general, for increasing robustness and generalizability, the feature set should to be extended.

The flow state refers to a state of mind that people experience when they act with total involvement.[38] Antecedents of flow are clear goals, unambiguous feedback and the challenge of the task meets the persons skills (skill-challenge balance).[39] Characteristics of being in flow are a strong focus on the task and a feeling of control, the merge of action and awareness, a loss of self-consciousness and a transformation of time.[39] Flow theory is tightly connected to attention and information processing theory because attention plays a critical role in achieving flow as it determines what we perceive. Furthermore, attention is a necessary condition for subsequent mental processes and events for flow.[38] Therefore, we argue that eye tracking can be a suitable technology to detect flow continuously and in real-time. So far, flow was already investigated using biosensors that rely on EEG and ECG technology.[40,41] However, despite the theoretical evidence, detecting the flow state using eye-tracking technology to the best of our knowledge was not evaluated so far.

In this study we continue the line of previous research by harnessing the power of eye-tracking data collected in the field to recognize cognitive user states. Consequently, we do not only focus on recognizing cognitive load but we also examine the recognition of flow using eye-tracking data. In addition, our research broadens the scope of cognitive state recognition by considering a wide range of tasks collected across several users.

## 2.2 Data collection methods: experience sampling method & ecological momentary assessment

The Experience Sampling Method (ESM), also known as Ecological Momentary Assessment (EMA), is a diary method for systematically obtaining self-reports from people about what they do, feel and think during activities in their daily lives.[18] While ESM primarily focuses on representativeness, EMA focuses more on momentariness of recorded survey data. However, there is no strict difference between the two methods.[16,42] Participants in ESM studies are typically given a pager, pen and paper, mobile device, or PC-based application that interrupts and displays a self-report questionnaire capturing current experiences at random points in their daily lives.[18] The collected self-reported data can be used to analyze affective and cognitive user states.[43–45] In addition, biosignal data collected along with the self-reported data can be used to detect changes in affective and cognitive user states.[18,46] A common approach is to analyze the biosignal data, such as ECG, EEG, or eye-tracking data, collected just prior to questionnaire administration in the context of the self-reported affective or cognitive state questionnaire responses.[44] Typically, the biosignal data from a given time window (e.g., 5 s, 30 s, 1 min, or 3 min) is aggregated to a specific metric, such as mean heart rate variability, mean alpha power, mean fixation duration, and so on.[41,47] By correlating this data with the affective or cognitive state labels, or by using this data along with the affective or cognitive state labels as input to an ML classifier, insights into the user's cognitive or affective states can be gained. A common approach is also to examine and compare multiple time windows of biosignal data for their influence on the explainability of the self-reported data.[47,48] This is typically

done to investigate empirically what is the best window size to capture the responses of triggers in the biosignal data that explain best the self-reported affective or cognitive state.

ESM has the advantage that self-report data can be collected in the natural environment, immediately during the experience and for a range of different experiences.[19] However, ESM also places a high burden on participants as they are interrupted from their ongoing activity several times a day.[44] Previous studies also show that participants experience fatigue during data collection, as the questionnaires in ESM studies are typically repetitive.[19] If burden and fatigue are too high, the risk of dropout increases, which may ultimately affect data quality or the comparability of data from different participants.[49] Therefore, low effort solutions such as collecting the self-reported data at the task, rather than switching to a smartphone or pen and paper to collect self-reported data, are key to reducing burdens. Furthermore, the combination of an ESM study and biosignals data collection can support researchers in accessing more data sources. However, there is a lack of guidelines and tools for conducting ESM studies in combination with biosignals, despite being more susceptible to external influences during data collection.

# 3 Field study

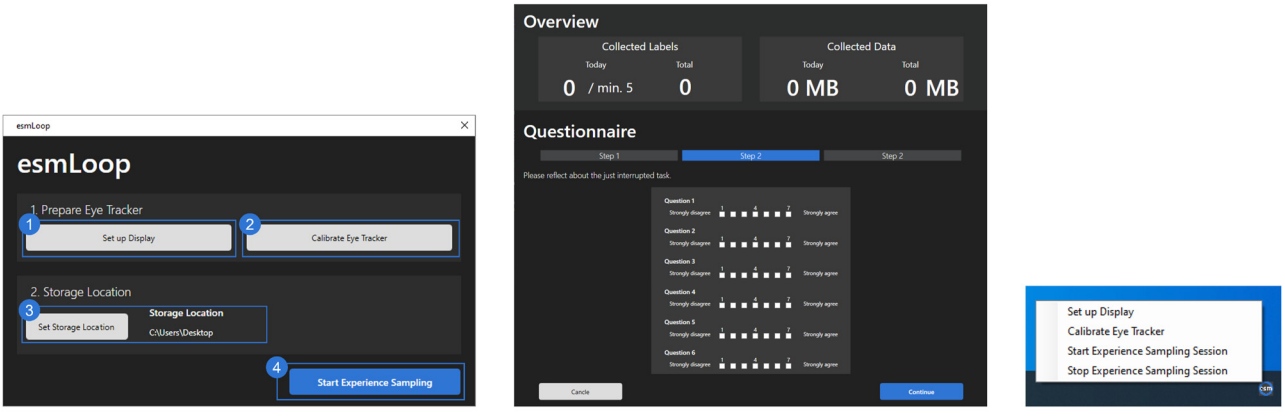
The goal of this field study is to explore the collection of cognitive state labels and eye-tracking data in the field and leverage the collected data to build models for two cognitive states, cognitive load and flow. As a foundation for our study we developed an experience sampling-based system called esmLoop. It supports collecting cognitive user state labels, eye-tracking and interaction data in the field.

## 3.1 The data collection system esmLoop

To conduct the field study, we developed esmLoop a PC-based desktop application that supports collecting data sets necessary for the development of supervised ML cognitive user state models. The user interface of esmLoop is depicted in Figure 1. In a first step, esmLoop guides the user through the process of setting up the eye tracker. When starting esmLoop, the start screen reminds the user to set up the display for the eye tracker (1). Subsequently, it requests calibrating the eye tracker as calibration is key to ensure high data quality collection (2). Furthermore, the user has to select the storage location for the recorded data to define where the data is stored (3). Currently esmLoop integrates the Tobii eye tracker 4C with the required research license that provides access to the Tobii Pro SDK. For the calibration of the Tobii 4C, we rely on the standard 7 points calibration provided by the Tobii 4C driver.

Once the user starts an experience sampling session by clicking the start button, data is recorded until the session is terminated by the user. The users have full control on starting and ending the recording of the data (4). Furthermore, the system can always be reached by the user through the icon on the task bar as shown in Figure 1c.

In terms of data collection, esmLoops records raw gaze data and pupil size during the experience sampling sessions. In addition, the title of the active window including a timestamp is recorded as well when the user swaps to another application. esmLoop issues a questionnaire at a random point in time every 20–60 min. Here, a message box pops up and asks whether the user is available to answer a short experience sampling questionnaire. If yes, the user is forwarded to the questionnaire window as demonstrated in Figure 1b. The questionnaire window is divided into two areas. The top area provides information about collected




**(a)** Start screen of esmLoop software

**(b)** Questionnaire screen of esmLoop Software

**(c)** Task Bar Menu

**Figure 1:** esmLoop user interface design.

labels and data. The user gets a transparent information about how many labels and mega bytes of data were collected on a specific day and overall. The bottom part shows the likert-scale based questions of the experience sampling questionnaire. In this study, we specifically asked participants adapted version the flow short scale and the NASA TLX questionnaire[32,50] on a 7 point likert scale. Furthermore, we also survey the currently interrupted task.

## 3.2 Study design

The study design went through the institutional review process in terms of ethics and data security and was approved prior to the study. This study focused on university students working on their final thesis projects. We selected thesis work because students continuously work on their thesis projects for several hours a day over several weeks, most phases of a thesis have to be completed at a computer, and it involves a variety but limited set of tasks. In addition, students experience higher levels of cognitive load during thesis work because the goal of the thesis is usually to work on a complex task or problem. In addition, thesis work typically also fulfills the requirements established by Nakamura and Csikszentmihalyi[39] of meeting the skill-challenge balance, having a clear goal, and experiencing unambiguous feedback about the progress of solving the tasks.

### 3.2.1 Participants

We recruited eight Bachelor and four Master students (total: 12 participants (4 female, 8 male)) with an average age of 25.03 years (SD = 3.15 years) who were invited through a university experimental lab panel and were working on their thesis project. In order to be eligible for participation, students had to work for at least 4 h per day on their thesis project for the duration of the study and they had to use their own PC or notebook for the data collection. To provide some flexibility for the students, they could select a minimum of 5 days during a time frame of 7 days for the data collection. Furthermore, we allowed participants to use their personal computers and collect data at any location (e.g., at home, library, student room etc.) that they would also work at under normal circumstance in order to increase external validity. Participants received 100€ as a compensation for study participation.

Later, we excluded one participant (female, P12) due to technical issues with the eye-tracking software driver during the experiment and continued with remaining 11 students. Two of the participants had previous experience with eye-tracking technology. All participants had a normal or corrected to normal vision except one participant who had one eye and a second glass eye. Monitor setups varied from single monitor to dual monitor setups with a screen resolutions between 1920 × 1080 pixels to 3000 × 2000 pixels. If participants had a dual monitor setup, we required them to install the eye tracker on the main monitor based on their own evaluation.

### 3.2.2 Procedure

To execute the study, we first invited all participants to an introduction workshop to introduce them to the study design, cognitive load and flow measurement, as well as eye-tracking technology. In this workshop, participants installed the required software esmLoop and the eye tracker jointly with the experimenter on their own private computer. To validate the correctness of the setup, they examplarily ran through the daily procedure of an experience sampling session within the esmLoop software including the setup and calibration of the eye tracker. After the introduction workshop, participants were ready for the actual field study and could start with the first experience sampling session.

The total duration of all experience sampling sessions per day were required to be higher than 4 h. They were allowed to be split into several session but each session had to be at least 60 min in order to be able to reach flow during that session. We decided for the 60 min minimum as many students follow a time boxing technique like Pomodoro technique,[51] meaning that after 1 h of focused work they take a 5–10 min break.

Participants were asked to calibrate the eye tracker and check if the data recording works before they started the session as shown in Figure 2. During the session, they were interrupted every 20–60 min at a random point of time by the experience sampling questionnaire. If the questionnaire prompt popped up at an inconvenient moment, e.g., during a video meeting, they were also allowed to postpone the questionnaire. By postponing the questionnaire, a new 20–60 min cycle was started. At the end of an experience sampling session, participants had to terminate the session in the esmLoop system. Once they collect more than 4 h of data on that day and finished the last session, participants had to upload the recorded data to a cloud drive in order to share the data for analysis. This procedure was repeated during minimum 5 days of the 7 days study time frame. At the end of the study, participants joined a final interview which was conducted in a semi-structured format and took 30 min. In this part, we asked interview questions about the experiences with the esmLoop system during data collection and labeling, about how they experience cognitive load and

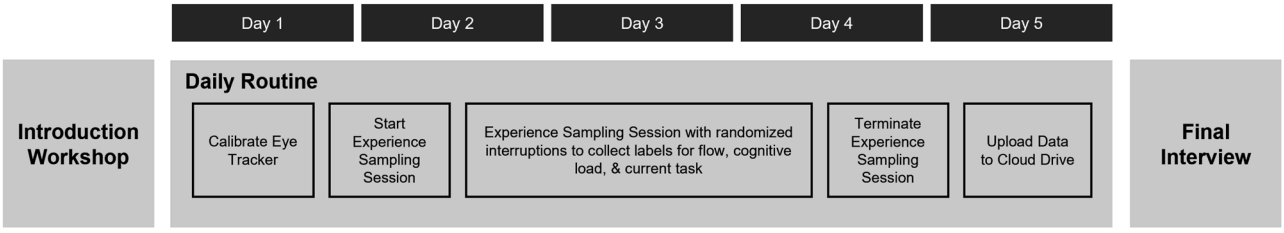



**Figure 2:** Procedure of study.

flow during the study, and about their perceived privacy and experiences with the eye tracker.

## 3.3 Data processing & modeling

In Figure 3 we visualized the data processing steps that were undertaken to develop and evaluate the cognitive state models. The eye-tracking data collected by the esmLoop software was pre-processed to extract fixations and saccades using the Pygaze Analyzer by Dalmaijer et al.[52]

First, we filtered the raw gaze data for gaze points that were marked as valid for both eyes and on the screen. Then, we calculated the average of both left and right eye as the Pygaze Analyzer only takes one $X$- & $Y$-coordinate as input for saccade and fixation calculation. For the participant with one glass eye, we considered only valid data from the real eye and skipped the averaging step. Next, we converted the normalized $X$- and $Y$-coordinates from the Tobii Pro SDK to pixel coordinates based on a $1920 \times 1080$ screen resolution to treat recorded data on different screens the same. Due to the frequency of the Tobii Eye Tracker 4C we set the minimum duration threshold for a fixation to 50 ms. As we experienced several outliers regarding fixation duration and saccade duration, we applied outlier detection (IQR $> 1.5$) and removed the detected fixation and saccade outliers. In addition, we decided to normalize the pupil size per session using min-max normalization to account for varying pupil size between sessions as pupil size is depending heavily on the light condition of the environment. After calculating the fixations and saccades, we extracted the features based on the fixations, saccades, and pupil size of different window sizes (1, 2, 3, 4, 5, 10, 20 min) before the questionnaire was issued. For example, for the one-minute window size, we only considered the eye-tracking data collected 1 min before the software administered the questionnaire and calculated the features using the eye-tracking data during that time window. Exploring multiple window sizes is a common approach in the development of eye-based models,[27,48] as eye-tracking data has a spatial and temporal dimension. We calculated the following features: fixation count, fixations per second, total fixation duration, mean fixation duration, saccade count, saccades per second, total saccade duration, saccade amplitude, saccade velocity, saccade acceleration, saccade absolute angle, saccade relative angle, saccade-fixation ration, fixation-saccade ratio, pupil diameter. We also calculated statistical features like mean, median, standard deviation, minimum, maximum, skew, and kurtosis where applicable. In total, we calculated 52 features (9 fixation-based, 36 saccade-based, 5 pupil-based and 2 ratio-based) as input for the model development. To define the cognitive load and flow state labels of participants, we averaged their answers on the 7 point likert scale regarding the 5 item NASA TLX questionnaire (excluded physical effort) and the first 10 items of the Flow Short Scale. We excluded the physical effort dimension of NASA TLX questionnaire as thesis writing does not vary in terms of physical effort. For the flow labels we focused on the

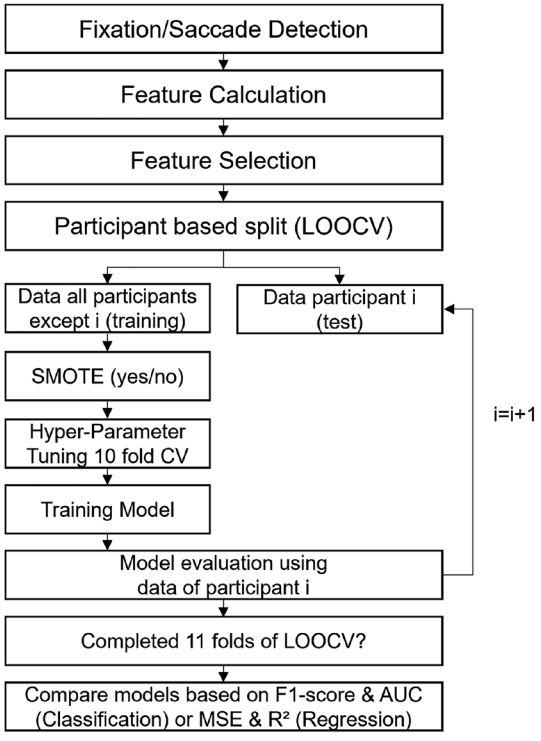



**Figure 3:** Data processing & modelling procedure.

first 10 items of the Flow Short Scale as these represent the flow experience while the later three questions focus on concerns.

The processing and modeling python scripts including the aggregated data frames of all above mentioned eye-tracking features and cognitive state labels can be downloaded here.[53] The following steps can also be found in the python scripts for the classification and regression. We decided to follow a binary classification approach as it was done in previous eye-tracking studies (e.g.[23,27,54,55]) and also a regression-based approach as likert-scale data is suitable for regression models. The following paragraphs are descriptions of the conducted modeling steps in the python scripts.

As the first step we applied the SelectFromModel function of the sklearn package [1] to reduce the set of features and therefore tried to avoid over-fitting of the models. To increase generalizability and robustness to unseen data we decided for a leave-one-out cross validation (LOOCV). This means that we considered data of 10 participants for training and fitting the model and then later evaluated the model's performance based on the data of the 11th participant. This procedure was repeated 11 times so that each participant's data served once as a test data set. As the next step, we splitted the data into a training and test data set. Since the cognitive load and flow labels were not evenly distributed across the binary classes of high and low cognitive load and flow and no flow, we decided to also test oversampling minority classes using the Synthetic Minority Oversampling Technique (SMOTE) function of the imblearn package[2] for the classification models. Then, we conducted hyper-parameter tuning using grid search with 10-fold cross validation on the data set of 10 participants. After model fitting, the performance of the model was evaluated by using the data set of the 11th participant. We developed the classification models using XGBoost (XGB), Random Forest (RF), CART Decision Tree (DT) algorithms of the sklearn package and compared them to the baseline majority class model (Base) for various window sizes (1, 2, 3, 4, 5, 10, 20 min). To calculate the overall performance of the classification models, the F1-Score and Area Under Curve (AUC) scores of all 11 folds of the LOOCV were averaged. We decided to use these two metrics for evaluation as they reflect the accuracy, generalizability and robustness of a model. For the regression models we used Linear Regression (LR), Decision Tree Regressor (DTR), Random Forest Regressor (RFR) algorithms of the sklearn package and for XGB Regressor (XGBR) the algorithm of the XGBoost package[3] and compared it to a baseline model always predicting the mean of the cognitive state labels of the test user. The overall performance of the regression models was evaluated by calculating the average of all mean square error (MSE) and the mean $R^2$ across all 11 folds of the LOOCV. This procedure was repeated 10 further times until all participants' data served once as a test data set.

# 4 Results

## 4.1 Descriptive data

### 4.1.1 Recorded eye-tracking data

In total, more than 250 h or 15,000 min of data were collected from the 11 participants in our field study. As shown in the Table 1, we were able to collect approximately 186.84 h of valid eye-tracking data. The Tobii SDK provides a Boolean value for each eye separately, indicating whether the eye tracker was able to correctly calculate the gaze point for that eye. In total, 59.54 % of the time (151.23 h) the participants collected eye-tracking data that was marked as valid for both eyes (see Table 1) and 73.56 % of the time at least one eye was marked as valid, which is also in line with a previous longitudinal eye-tracking field study.[56] It should be noted that invalid eye-tracking data can occur when the participant is not present, looking at the screen, looking at another screen, or due to technical problems. As this is a field study, it must be emphasized that participants may have left the computer for short breaks during data collection.

In Tables 2 and 3, we demonstrate descriptive data for selected eye-tracking data based features of flow/no flow and high/low cognitive load labels. We report fixation duration (total fixation duration during the time window divided by window size), fixation count per second, saccade duration (total saccade duration during the time window divided

**Table 1:** Collected valid eye tracking data.

| | Both eyes valid | At least one eye valid | Total |
|---|---|---|---|
| Duration | 151.23 h | 186.84 h | 254.01 h |
| Percentage | 59.54 % | 73.56 % | 100 % |

1 https://scikit-learn.org/

2 https://pypi.org/project/imblearn/

3 https://pypi.org/project/xgboost/

**Table 2:** Descriptive eye tracking data for flow.

| Window size | Fixation dur. [flow] | Fixation dur. [no flow] | Fixation count [flow] | Fixation count [no flow] | Saccade dur. [flow] | Saccade dur. [no flow] | Saccade count [flow] | Saccade count [no flow] | Pupil size mean [flow] | Pupil size mean [no flow] |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 45 % | 41 % | 2.59 | 2.46 | 50 % | 48 % | 1.55 | 1.29 | 0.58 | 0.58 |
| 2 | 44 % | 40 % | 2.54 | 2.39 | 50 % | 45 % | 1.55 | 1.31 | 0.58 | 0.59 |
| 3 | 43 % | 39 % | 2.50 | 2.31 | 48 % | 43 % | 1.53 | 1.32 | 0.57 | 0.59 |
| 4 | 42 % | 39 % | 2.43 | 2.30 | 46 % | 42 % | 1.48 | 1.30 | 0.56 | 0.58 |
| 5 | 42 % | 39 % | 2.41 | 2.31 | 44 % | 41 % | 1.46 | 1.29 | 0.55 | 0.57 |
| 10 | 42 % | 39 % | 2.41 | 2.30 | 41 % | 40 % | 1.38 | 1.27 | 0.54 | 0.54 |
| 20 | 42 % | 39 % | 2.40 | 2.33 | 39 % | 39 % | 1.32 | 1.23 | 0.53 | 0.53 |

**Table 3:** Descriptive eye tracking data for cognitive load.

| Window size | Fixation dur [high] | Fixation duration [low] | Fixation count [high] | Fixation count [low] | Saccade dur [high] | Saccade dur [low] | Saccade count [high] | Saccade count [low] | Pupil size mean [high] | Pupil size mean [low] |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 46 % | 40 % | 2.64 | 2.40 | 52 % | 46 % | 1.49 | 1.37 | 0.59 | 0.57 |
| 2 | 46 % | 38 % | 2.67 | 2.22 | 51 % | 44 % | 1.58 | 1.27 | 0.59 | 0.57 |
| 3 | 45 % | 38 % | 2.61 | 2.17 | 49 % | 42 % | 1.54 | 1.31 | 0.58 | 0.57 |
| 4 | 44 % | 37 % | 2.56 | 2.13 | 47 % | 40 % | 1.52 | 1.26 | 0.57 | 0.56 |
| 5 | 44 % | 37 % | 2.56 | 2.11 | 45 % | 39 % | 1.50 | 1.24 | 0.57 | 0.55 |
| 10 | 44 % | 36 % | 2.57 | 2.09 | 44 % | 36 % | 1.41 | 1.23 | 0.55 | 0.53 |
| 20 | 44 % | 37 % | 2.56 | 2.13 | 42 % | 35 % | 1.33 | 1.22 | 0.54 | 0.51 |

by window size), saccade count per second, and mean normalized pupil size to account for the influence of window size. It can be seen that the average of almost all fixation- and saccade-based metrics is higher for the flow/high cognitive load labels than for the no flow/low cognitive load labels. Moreover, the larger the time window, the lower the differences in the features between the flow/high cognitive load labels than for the no flow/low cognitive load labels.

#### 4.1.2 Recorded label data

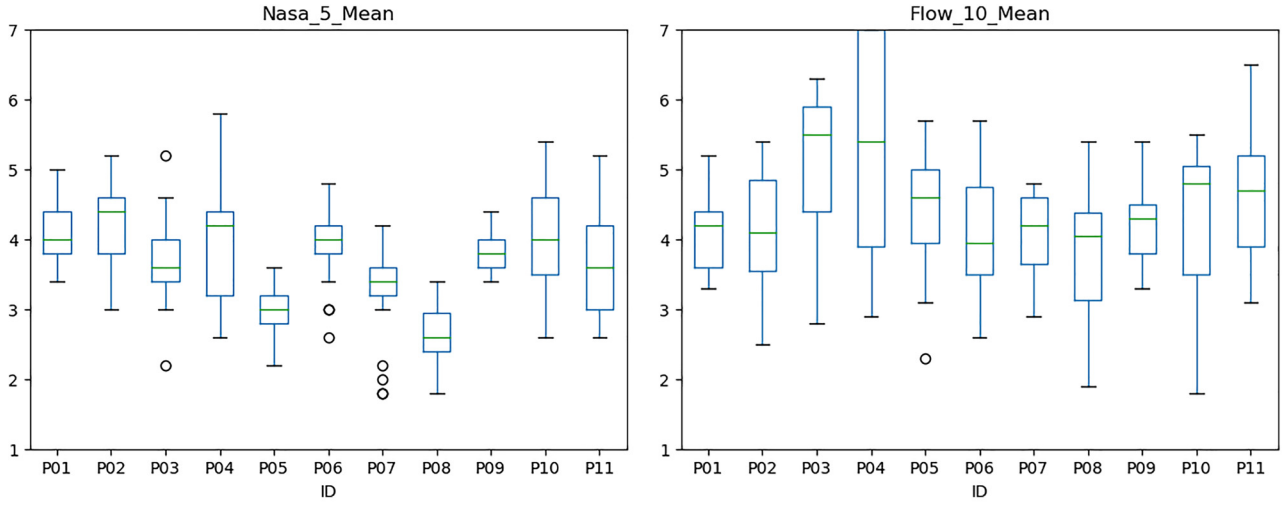
During this study, participants completed 293 experience sampling questionnaires about their current cognitive load, flow state, and current task. We observed a compliance rate of 98.75 %, which means that 98.75 % of all questionnaires distributed were actually answered by the participants, and only a small fraction of 1.25 % of the questionnaires were postponed. However, we did not incentivize the number of answered questionnaires.

The averaged responses of the cognitive load (NASA TLX) and flow state (Flow Short Scale) questionnaires are shown in Figure 4a and b. From these visualizations we can see that the participants experienced cognitive load and flow state differently during the study.

We normalized the averaged responses using a min–max normalization since the final interview results confirm that all participants experienced both high and low cognitive load as well as flow and no flow during the study. The normalized cognitive load and flow state distribution can be seen in Figure 5a and b. Finally, we considered a label as high cognitive load or flow if the normalized mean was greater than 0.5 (visualized by the red line) and otherwise as low cognitive load or no flow.

In Table 4 you can see the distribution of cognitive load and flow labels before and after normalization. Before normalization, we considered a label as high cognitive load or flow if the averaged questionnaire answers were greater than 4. We can see that the label distribution became closer to a 50:50 distribution due to normalization.

In addition, we examined the tasks reported in the experience sampling questionnaires to explore the tasks that users were engaged in during the experiment. Figure 6 visualizes the number of tasks that participants were working on when they were interrupted. Since we recruited participants who were working on their thesis project, the two most frequently reported tasks were writing a text (94 times) and literature research (87 times). Both tasks account for more than half of all recorded task labels. Other tasks




**(a)** Mean cognitive load labels by participant

**(b)** Mean flow labels by participant

**Figure 4:** Mean cognitive state labels.

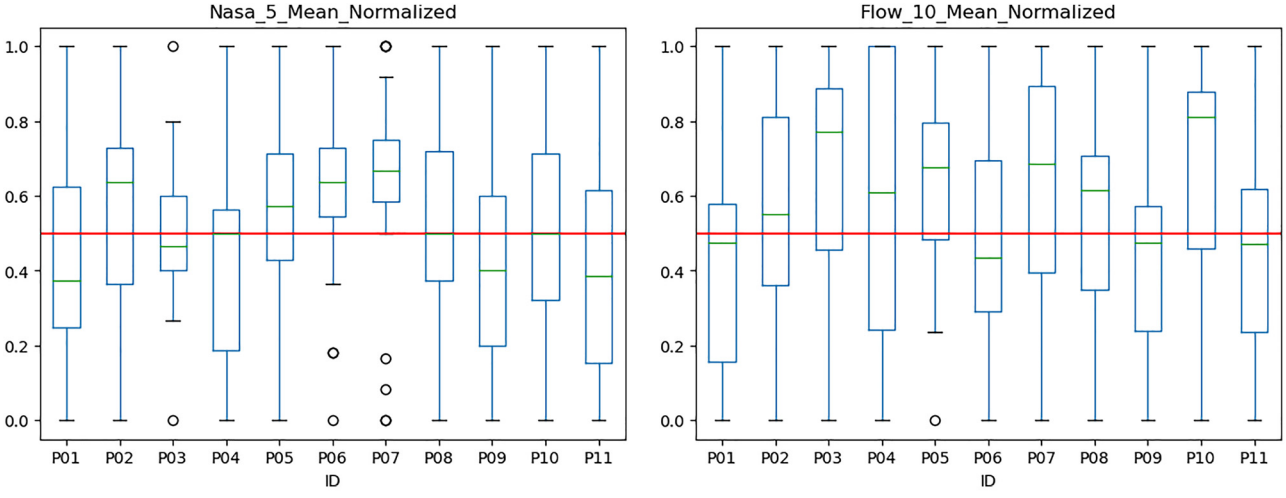
were other tasks (30 times), proofreading (19 times), and creative tasks (19 times).

#### 4.1.3 Recorded interaction data

To provide further context to the collected labels, we also recorded the duration that an application was the active window on the screen during the entire experience sampling session. Figure 7 shows all of the applications we tracked and the total amount of time each application was the active window on the screen. The analysis shows that the most used application was the internet browser (87.60 h). The second and third most used applications were Word (58.40 h) and Overleaf (38.38 h), reflecting the writing task of thesis projects. The "Other" block (29.29 h) represents all other applications that were not tracked to increase privacy and may not be related to thesis project work, such as Spotify.

## 4.2 Interview results

At the end of the study, each participant was interviewed for 30 min. After using the esmLoop for 5 days, 91 % of participants (10/11) rated their experience with the software and data collection process as positive, and they experienced only minor problems during the course of using the




**(a)** Normalized mean cognitive load labels by participant

**(b)** Normalized mean flow labels by participant

**Figure 5:** Normalized mean cognitive state labels.

**Table 4:** Label distribution of high/low cognitive load and flow/no flow before and after normalization.

| | High cognitive load | Low cognitive load | Flow | No flow |
|---|---|---|---|---|
| Not normalized | 37.9 % | 62.1 % | 65.6 % | 34.4 % |
| Normalized | 57.0 % | 43.0 % | 57.7 % | 42.3 % |

esmLoop. Only one participant rated the experience as moderate due to connectivity issues with the eye tracker. In addition, P04, P06, and P08 particularly emphasized that the introductory workshop was helpful and supported them in familiarizing themselves with the study and the eye-tracking technology.

Regarding the **labeling task**, 36 % of the participants (4/11) stated that filling out the questionnaires was sometimes interrupting, while 18 % of the participants (2/11) did not feel interrupted in their work. For example, P03 said that the esmLoop questionnaire schedule fitted to her schedule as she was following the Pomodoro approach with 50 min focused work and 10 min break. P01 and P05 reported that their usual break schedule was partially affected by esmLoop as they wanted to wait until the next questionnaire pops up to finally have a break. Furthermore, four interviewees stated that the report of the number of collected labels (see Figure 1b) was useful to see a progress in the data collection process and stay motivated.

Regarding **cognitive load**, 36 % of the participants (4/11) experience a high cognitive load several times during the study. 27 % of interviewees stated that most of the time they operated on a medium level of cognitive load. Participants defined cognitive load based on different dimensions, like time pressure, focus, stress or a challenging tasks. However, five out of 11 interviewees regarded cognitive load as a spectrum while one interview would define it as binary. Moreover, P09 stated that *"for a good flow you need a high mental workload"* and P06 reported that *"when you have a task that requires a high mental workload, you definitely get into the flow state easily"*. All in all these two participants

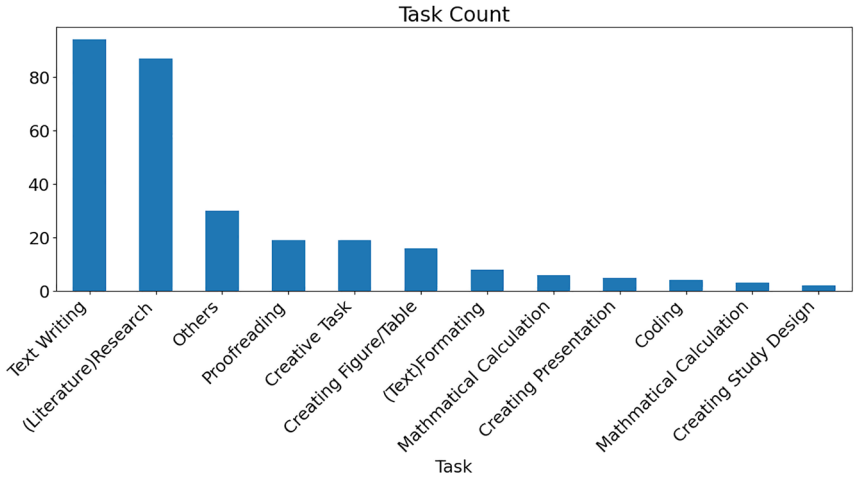



**Figure 6:** Overview of tasks that participants have been accomplishing when answering the questionnaire about their flow and cognitive load state.

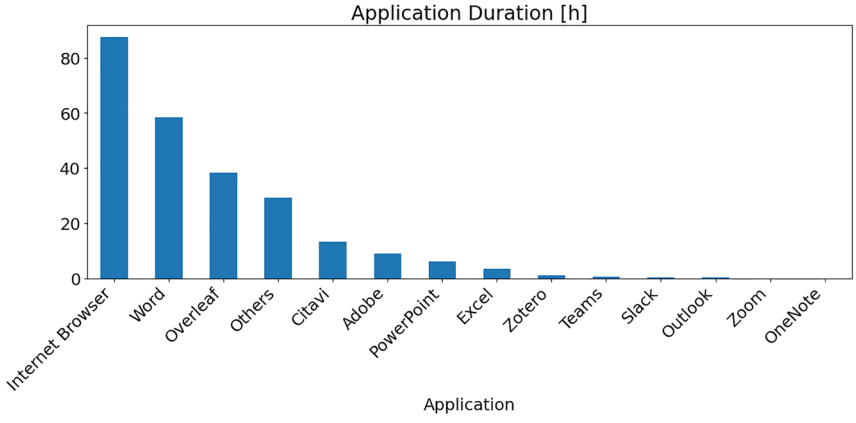



**Figure 7:** Applications.

related a high cognitive load to flow as either a requirement or a facilitator.

Regarding experiencing **flow** during the task, the interview results highlight that all participants experienced flow during this study differently. The number of times and intensity of experiencing flow varied across the study's participants. Two participants (P02, P04) stated that they experienced more flow during the study than they would have anticipated. On the other side, P05 disclosed that he was more often not in flow than in flow. Furthermore, five participants also associated flow with productivity. P01 stated *"I was already productive and noticed that I was productive and then when it was a very strong flow (…) then it was so that I didn't even think about it anymore"*. The timing of the questionnaire sometimes had an influence on their flow according to four participants as they were either interrupted in the middle of finishing a task or they waited for the next questionnaire to pop up.

In terms of **privacy and data protection**, 72 % of interviewees (8/11) had no concerns during the study. At the beginning three participants felt observed due to the eye tracker and recorded data but this feeling vanished over the course of the study. However, two of them reported that this effect got less over the course of the study as they got used to it and that it increased their focus on the thesis. Four interviewees also reported that the eye tracker did not disturb them. Regarding the **eye-tracking technology**, in general, many participants were surprised by the accuracy of the eye tracker. However, 82 % of participants (9/11) reported problems with the Tobii 4C eye tracker because it sometimes randomly disconnected itself from the computer while using it for a longer time. This created some frustration at the participants site as P03 reported that *"the recording has been running for three quarters of an hour and then an error message comes out of nowhere, that's frustrating"*. Furthermore, two participants feared that the red infrared light sources of the eye tracker would bother them which eventually was not the case.

## 4.3 Cognitive state models

In chapter 3 we described the data processing & modelling procedure in detail which we followed to develop the cognitive state models for cognitive load and flow using the collected field data.

### 4.3.1 Models considering all labels

First, we developed the cognitive state models using all the label data collected during the study to investigate the feasibility of generalizable and accurate eye-based cognitive state models using field data. We used the SelectFromModel function of the sklearn package to reduce the feature set to the most important features for each algorithm separately. After selecting the features, we trained with and without applying SMOTE and evaluated the models using LOOCV.

#### 4.3.1.1 Classification models for all tasks

For the classification models, we evaluated the performance of the models based on the average F1-Score and AUC of the LOOCV and the results can be found in Table 5. If we compare the performance of the classifiers without SMOTE/oversampling with the baseline classifier, we can see that for flow only, the Random Forest classifier performed as well as the baseline classifier for both F1-Score ($F1-Score = 72.80$) and AUC ($AUC = 0.5$) for a window size of 1 min. However, a closer look at the confusion matrices of each LOOCV step shows that the classifier always predicted the negative class, i.e., no flow, in 11 out of 11 cross-validation steps. This shows that the classifier has little discriminative power and therefore does not show generalizability. Considering SMOTE/oversampling for the evaluation of the flow classifier, a Decision Tree classifier with a window size of 10 min performed best in terms of F1-Score ($F1-Score = 62.12$, $AUC = 0.542$), while a Random Forest classifier with a window size of 10 min performed best in terms of AUC ($AUC = 0.586$). However, none of the classifiers outperformed the baseline classifier for both metrics using SMOTE/oversampling.

Similar results can be observed for the cognitive load classifier. Without SMOTE/oversampling, the best performing classifier in terms of F1-Score is based on Decision Trees and a window size of 3 min ($F1-Score = 71.09$, $AUC = 0.517$) and outperformed the baseline model by a small margin, but exhibited a low discriminative power slightly above 0.5. In terms of AUC an XGBoost based classifier and a window size of 3 min is performing best ($AUC = 0.566$) but did not outperform the baseline model in terms of F1-Score. Using SMOTE/oversampling, the best performing classifier in terms of F1-Score is based on Decision Trees and a window size of 2 min ($F1-Score = 65.83$) and in terms of AUC, a Random Forest based classifier with a window size of 20 min performs best ($AUC = 0.579$). Overall, none of the classifiers performed significantly better than the baseline classifier for both metrics. Furthermore, none of the window sizes significantly outperformed the other window sizes.

#### 4.3.1.2 Regression models for all tasks

For the regression models, we evaluated the performance of the models based on the average MSE and $R^2$ of the LOOCV

**Table 5:** Classifier performance for all window sizes (1, 2, 3, 4, 5, 10, 20 min) and classifiers (DT, RF, XGB, Base) including labels from all tasks. Bold marked results of classifiers outperformed the baseline classifier of the corresponding window size in terms of both metrics.

| Window size | Classifier | Flow | | Flow [oversampled] | | Cognitive load | | Cognitive load [oversampled] | |
|---|---|---|---|---|---|---|---|---|---|
| | | F1-Score [%] | AUC | F1-Score [%] | AUC | F1-Score [%] | AUC | F1-Score [%] | AUC |
| 1 | DT | 72.58 | 0.498 | 60.40 | 0.471 | 67.73 | 0.506 | 57.51 | 0.491 |
| 1 | RF | 72.80 | 0.500 | 51.13 | 0.470 | 68.06 | 0.498 | 56.48 | 0.550 |
| 1 | XGB | 66.17 | 0.495 | 51.51 | 0.478 | 63.92 | 0.556 | 57.22 | 0.530 |
| 1 | Base | 72.80 | 0.500 | 66.67 | 0.500 | 69.51 | 0.500 | 66.67 | 0.500 |
| 2 | DT | 67.67 | 0.496 | 62.12 | 0.490 | 69.77 | 0.528 | 65.83 | 0.563 |
| 2 | RF | 71.22 | 0.506 | 55.66 | 0.499 | 70.28 | 0.558 | 57.88 | 0.565 |
| 2 | XGB | 60.40 | 0.493 | 54.16 | 0.505 | 66.20 | 0.561 | 57.34 | 0.577 |
| 2 | Base | 71.78 | 0.500 | 66.67 | 0.500 | 70.33 | 0.500 | 66.67 | 0.500 |
| 3 | DT | 71.08 | 0.509 | 59.32 | 0.542 | **71.09** | **0.517** | 61.67 | 0.497 |
| 3 | RF | 70.41 | 0.545 | 52.07 | 0.481 | 64.97 | 0.521 | 56.60 | 0.524 |
| 3 | XGB | 64.58 | 0.511 | 51.75 | 0.480 | 65.82 | 0.566 | 58.16 | 0.564 |
| 3 | Base | 71.13 | 0.500 | 66.67 | 0.500 | 70.75 | 0.500 | 66.67 | 0.500 |
| 4 | DT | 70.86 | 0.496 | 58.57 | 0.500 | 66.97 | 0.497 | 56.52 | 0.503 |
| 4 | RF | 69.82 | 0.523 | 57.40 | 0.555 | 69.15 | 0.534 | 46.64 | 0.467 |
| 4 | XGB | 65.70 | 0.518 | 56.06 | 0.526 | 61.56 | 0.499 | 54.81 | 0.513 |
| 4 | Base | 71.44 | 0.500 | 66.67 | 0.500 | 69.90 | 0.500 | 66.67 | 0.500 |
| 5 | DT | 66.27 | 0.495 | 55.36 | 0.462 | 68.10 | 0.493 | 60.57 | 0.483 |
| 5 | RF | 69.46 | 0.523 | 53.39 | 0.537 | 67.23 | 0.523 | 53.35 | 0.521 |
| 5 | XGB | 63.36 | 0.502 | 56.17 | 0.546 | 59.71 | 0.511 | 55.65 | 0.539 |
| 5 | Base | 71.53 | 0.500 | 66.67 | 0.500 | 70.11 | 0.500 | 66.67 | 0.500 |
| 10 | DT | 66.68 | 0.528 | 62.12 | 0.542 | 69.05 | 0.530 | 63.98 | 0.539 |
| 10 | RF | 69.77 | 0.519 | 61.12 | 0.586 | 65.27 | 0.502 | 53.72 | 0.520 |
| 10 | XGB | 67.79 | 0.588 | 55.02 | 0.554 | 59.01 | 0.505 | 50.59 | 0.478 |
| 10 | Base | 71.55 | 0.500 | 66.67 | 0.500 | 70.28 | 0.500 | 66.67 | 0.500 |
| 20 | DT | 71.26 | 0.505 | 51.29 | 0.517 | 69.34 | 0.515 | 65.44 | 0.507 |
| 20 | RF | 70.59 | 0.511 | 53.08 | 0.531 | 67.77 | 0.515 | 58.35 | 0.579 |
| 20 | XGB | 64.71 | 0.510 | 52.81 | 0.514 | 57.12 | 0.525 | 53.68 | 0.517 |
| 20 | Base | 71.55 | 0.500 | 66.67 | 0.500 | 70.28 | 0.500 | 66.67 | 0.500 |

and the results can be found in Table 6. Note that $R^2$ can become negative if the residual sum of squares is very large, indicating a poor fit of the model.

Comparing the performance of the regressions without label normalization to the baseline model, we see that for flow, the Decision Tree Regressor with a window size of 3 min performed best in terms of MSE (MSE = 0.969), while for $R^2$, Decision Tree Regressor with a window size of 2 min performed best ($R^2 = 0.047$). This Decision Tree regressor also outperformed the basline model by a small margin regarding the MSE (MSE = 1.044), making it slightly superior to the baseline model. However, an $R^2$ of 0.047 indicates that the model is performing relatively poorly on unseen data, as only 4.7 % of the variance in the dependent variable is explained by the independent variables included in the model. Applying normalization to the flow labels, the Random Forest Regressor for a window size of 3 min performed best regarding the MSE (MSE = 0.084) and also outperformed the baseline model by a small margin, while a Random Forest Regressor for a 20 min window size performed best for the $R^2$ score ($R^2 = 0.168$). This Random Forest Regressor model also performed as well as the baseline model in terms of the MSE (MSE = 0.088). Overall, none of the models for flow significantly outperformed the baseline model while also having a good predictive power.

For cognitive load, none of the regression models outperformed the baseline model in terms of MSE and achieved an $R^2$ score >0. The best performing model without normalization was a Random Forest Regressor for the

**Table 6:** Regression-based model performance for all window sizes (1, 2, 3, 4, 5, 10, 20 min) and regression based models (LR, DTR, RFR, XGBR, Base) including labels from all tasks. Bold marked results of models outperformed the baseline model of the corresponding window size in terms of MSE and had a $R^2 > 0$.

| Window size | Classifier | Flow | | Flow [normalized] | | Cognitive load | | Cognitive load [normalized] | |
|---|---|---|---|---|---|---|---|---|---|
| | | MSE | $R^2$ | MSE | $R^2$ | MSE | $R^2$ | MSE | $R^2$ |
| 1 | LR | 1.100 | −0.269 | 0.098 | −0.379 | 0.646 | −1.555 | 0.079 | −0.207 |
| 1 | DTR | 1.133 | −0.482 | 0.096 | −0.118 | 0.638 | −1.626 | 0.068 | −0.044 |
| 1 | RFR | 1.101 | −0.162 | 0.091 | −0.103 | 0.595 | −1.515 | 0.068 | −0.048 |
| 1 | XGBR | 1.349 | −0.433 | 0.111 | −0.224 | 0.743 | −2.230 | 0.079 | −0.201 |
| 1 | Base | 1.050 | −0.079 | 0.090 | −0.117 | 0.588 | −1.421 | 0.071 | −0.094 |
| 2 | LR | 1.086 | −0.033 | 0.094 | −0.139 | 0.605 | −1.377 | 0.069 | −0.102 |
| 2 | DTR | **1.044** | **0.047** | 0.096 | 0.115 | 0.626 | −1.510 | 0.070 | −0.126 |
| 2 | RFR | 1.093 | −0.019 | 0.091 | −0.001 | 0.600 | −1.324 | 0.066 | −0.063 |
| 2 | XGBR | 1.093 | −0.154 | 0.096 | 0.039 | 0.658 | −1.531 | 0.078 | −0.238 |
| 2 | Base | 1.051 | −0.046 | 0.090 | −0.155 | 0.584 | −1.322 | 0.068 | −0.081 |
| 3 | LR | 1.068 | −0.058 | 0.091 | −0.079 | 0.601 | −1.379 | 0.069 | −0.103 |
| 3 | DTR | 0.969 | −0.185 | 0.093 | −0.020 | 0.605 | −1.463 | 0.067 | −0.075 |
| 3 | RFR | 1.011 | −0.388 | 0.084 | −0.002 | 0.639 | −1.560 | 0.064 | −0.035 |
| 3 | XGBR | 1.300 | −1.155 | 0.104 | −0.244 | 0.665 | −1.690 | 0.081 | −0.301 |
| 3 | Base | 1.043 | −0.083 | 0.088 | −0.124 | 0.580 | −1.318 | 0.067 | −0.080 |
| 4 | LR | 1.034 | −0.125 | 0.091 | −0.110 | 0.589 | −1.283 | 0.068 | −0.107 |
| 4 | DTR | 0.995 | −0.412 | 0.095 | −0.225 | 0.610 | −1.487 | 0.065 | −0.065 |
| 4 | RFR | 1.050 | −0.496 | 0.085 | −0.107 | 0.613 | −1.423 | 0.065 | −0.079 |
| 4 | XGBR | 1.145 | −0.998 | 0.100 | −0.651 | 0.646 | −1.566 | 0.076 | −0.249 |
| 4 | Base | 1.017 | −0.086 | 0.087 | −0.125 | 0.567 | −1.217 | 0.065 | −0.068 |
| 5 | LR | 1.039 | −0.165 | 0.094 | −0.070 | 0.582 | −1.258 | 0.068 | −0.119 |
| 5 | DTR | 1.055 | −0.367 | 0.095 | −0.037 | 0.624 | −1.316 | 0.072 | −0.189 |
| 5 | RFR | 1.033 | −0.564 | 0.089 | 0.029 | 0.639 | −1.550 | 0.066 | −0.087 |
| 5 | XGBR | 0.983 | −0.631 | 0.095 | −0.343 | 0.651 | −1.561 | 0.082 | −0.382 |
| 5 | Base | 1.013 | −0.084 | 0.087 | −0.127 | 0.567 | −1.221 | 0.065 | −0.067 |
| 10 | LR | 1.077 | −0.218 | 0.129 | −0.357 | 0.593 | −1.287 | 0.065 | −0.084 |
| 10 | DTR | 1.016 | −0.317 | 0.091 | −0.013 | 0.575 | −1.235 | 0.063 | −0.055 |
| 10 | RFR | 1.037 | −0.327 | 0.088 | −0.005 | 0.551 | −1.207 | 0.063 | −0.077 |
| 10 | XGBR | 1.280 | −0.748 | 0.106 | −0.300 | 0.667 | −1.619 | 0.069 | −0.203 |
| 10 | Base | 1.000 | −0.084 | 0.088 | −0.129 | 0.562 | −1.225 | 0.064 | −0.067 |
| 20 | LR | 1.137 | −0.124 | 0.094 | −0.119 | 0.629 | −1.376 | 0.065 | −0.088 |
| 20 | DTR | 1.093 | −0.341 | 0.088 | −0.027 | 0.549 | −1.471 | 0.068 | −0.147 |
| 20 | RFR | 1.045 | −0.388 | 0.088 | 0.168 | 0.488 | −0.868 | 0.063 | −0.068 |
| 20 | XGBR | 1.031 | −0.219 | 0.089 | −0.096 | 0.531 | −1.620 | 0.073 | −0.234 |
| 20 | Base | 1.000 | −0.084 | 0.088 | −0.129 | 0.562 | −1.225 | 0.064 | −0.067 |

20 min window size when considering MSE (MSE = 0.488) and $R^2$ ($R^2 = -0.868$). This model actually outperformed the baseline model, but the $R^2$ value was negative, indicating a high residual sum of squares and low model fit. When normalizing the cognitive load labels, a Decision Tree Regressor and a Random Forest Regressor for the 10 min window size, and a XGBoot Regressor for the 20 min window size outperformed the baseline model in terms of MSE (MSE = 0.063). However, none of the regression models for normalized cognitive state labels could achieve a positive $R^2$ highlighting that none of the models had a good model fit.

Overall, the results for both the eye-based cognitive state classification and the regression models trained on all label data show that we have not been successful in building models that are generalizable across tasks and participants. None of the developed classification and regression models significantly outperformed the baseline models. Only one of the developed classifiers slightly outperformed the baseline

classifier in terms of F1-Score and AUC, and only one regression model slightly outperformed the baseline model while showing poor variance explainability.

### 4.3.2 Models considering writing task labels

As shown in the task and application analysis, the tasks varied significantly during the field study and also between participants. Therefore, we filtered our dataset for the most frequently reported task "text writing" and followed the same procedure as described above to evaluate the performance of different classifiers trained with only 94 "text writing" labels. However, not all participants reported labels for the "text writing" task. Therefore, only P01, P02, P04, P05, P06, P09, P10 could be considered in the following section.

#### 4.3.2.1 Classification models for writing tasks

In order to apply SMOTE with a minimum of $k = 2$ $k$-neighbors, at least three labels of the minority class must be collected. Therefore, we could only include P1, P2, P4, P6 and P9 for the development of the flow classifier and P1, P2, P4, P9, P10 for the development of the cognitive load classifier. The results of the classification models using only "text writing" labels can be found in Table 7. For flow, without SMOTE/oversampling, XGBoost outperformed the baseline classifier for the window sizes of 1 min ($F1 - Score = 67.54$, $AUC = 0.618$) and 2 min ($F1 - Score = 74.38$, $AUC = 0.691$). Using SMOTE to balance the flow labels, XGBoost outperformed the baseline classifier for a window size of 2 min and was the best overall performer for both metrics ($F1 - Score = 71.83$, $AUC = 0.689$). In addition, a Decision Tree classifier for the 20 min window also outperformed

**Table 7:** Classifier performance for all window sizes (1, 2, 3, 4, 5, 10, 20 min) and classifiers (DT, RF, XGB, Base) including labels from writing tasks only. Bold marked results of classifiers outperformed the baseline classifier of the corresponding window size in terms of both metrics.

| Window size | Classifier | Flow writing | | Flow writing [oversampled] | | Cognitive load writing | | Cognitive load writing [oversampled] | |
|---|---|---|---|---|---|---|---|---|---|
| | | F1-Score [%] | AUC | F1-Score [%] | AUC | F1-Score [%] | AUC | F1-Score [%] | AUC |
| 1 | DT | 60.74 | 0.463 | 52.73 | 0.452 | 68.46 | 0.517 | 49.45 | 0.493 |
| 1 | RF | 64.83 | 0.491 | 60.57 | 0.500 | **70.96** | **0.690** | 52.04 | 0.626 |
| 1 | XGB | **67.54** | **0.618** | 60.33 | 0.537 | 41.51 | 0.513 | 49.42 | 0.563 |
| 1 | Base | 67.12 | 0.500 | 66.67 | 0.500 | 70.43 | 0.500 | 66.67 | 0.500 |
| 2 | DT | 68.77 | 0.582 | 66.11 | 0.595 | 69.31 | 0.547 | **66.84** | **0.588** |
| 2 | RF | 66.32 | 0.506 | 63.98 | 0.563 | 67.46 | 0.580 | 64.95 | 0.635 |
| 2 | XGB | **74.38** | **0.691** | **71.83** | **0.689** | 54.83 | 0.540 | 50.92 | 0.534 |
| 2 | Base | 66.38 | 0.500 | 66.67 | 0.500 | 71.26 | 0.500 | 66.67 | 0.500 |
| 3 | DT | 65.76 | 0.521 | 56.35 | 0.489 | **74.64** | **0.635** | 56.30 | 0.510 |
| 3 | RF | 62.94 | 0.498 | 58.67 | 0.503 | 65.72 | 0.605 | 62.95 | 0.616 |
| 3 | XGB | 52.58 | 0.490 | 57.07 | 0.536 | 70.40 | 0.678 | 59.64 | 0.626 |
| 3 | Base | 65.80 | 0.500 | 66.67 | 0.500 | 71.63 | 0.500 | 66.67 | 0.500 |
| 4 | DT | 62.34 | 0.517 | 61.23 | 0.450 | **70.80** | **0.603** | 63.99 | 0.581 |
| 4 | RF | 61.53 | 0.512 | 65.89 | 0.607 | 67.46 | 0.550 | 59.77 | 0.633 |
| 4 | XGB | 55.66 | 0.524 | 60.70 | 0.550 | 60.56 | 0.604 | **70.53** | **0.711** |
| 4 | Base | 65.80 | 0.500 | 66.67 | 0.500 | 70.68 | 0.500 | 66.67 | 0.500 |
| 5 | DT | 65.80 | 0.500 | 58.47 | 0.515 | 64.65 | 0.520 | 63.29 | 0.553 |
| 5 | RF | 59.95 | 0.475 | 56.02 | 0.480 | 61.39 | 0.490 | 42.63 | 0.467 |
| 5 | XGB | 51.40 | 0.442 | 57.05 | 0.567 | 66.62 | 0.541 | 56.88 | 0.524 |
| 5 | Base | 65.80 | 0.500 | 66.67 | 0.500 | 71.46 | 0.500 | 66.67 | 0.500 |
| 10 | DT | 58.29 | 0.439 | 62.30 | 0.492 | 72.08 | 0.500 | 63.70 | 0.620 |
| 10 | RF | 60.72 | 0.469 | 59.29 | 0.532 | 62.01 | 0.545 | 55.95 | 0.573 |
| 10 | XGB | 54.38 | 0.456 | 60.90 | 0.523 | 65.25 | 0.553 | 59.77 | 0.529 |
| 10 | Base | 65.80 | 0.500 | 66.67 | 0.500 | 72.08 | 0.500 | 66.67 | 0.500 |
| 20 | DT | 64.90 | 0.491 | **70.00** | **0.571** | 68.01 | 0.477 | 54.09 | 0.427 |
| 20 | RF | 64.16 | 0.483 | 58.79 | 0.473 | 56.44 | 0.503 | 53.04 | 0.536 |
| 20 | XGB | 62.02 | 0.510 | 62.78 | 0.549 | 53.11 | 0.511 | 53.63 | 0.528 |
| 20 | Base | 65.80 | 0.500 | 66.67 | 0.500 | 72.08 | 0.500 | 66.67 | 0.500 |

the baseline classifier for both metrics ($F1-Score = 70.00$, $AUC = 0.571$), but performed slightly worse than the XGBoost classifier for the 2 min window. For cognitive load and without SMOTE/oversampling, a Decision Tree classifier produced the highest F1-Score ($F1-Score = 74.64$) and in terms of AUC a Random Forest classifier for a window size of 1 min ($AUC = 0.66$). Applying SMOTE/oversampling, a XGBoost classifier for a window size of 4 min performed best in terms of the F1-Score and AUC ($F1-Score = 70.53$, $AUC = 0.711$) while also outperforming the baseline classifier.

#### 4.3.2.2 Regression models for writing tasks

The results of the regression models using only "text writing" labels can be found in Table 8. Without normalizing the flow labels, an XGBoost Regressor for a window size of 2 min performed best in terms of the $R^2$ value ($R^2 = 0.229$), while also outperforming the baseline regression

**Table 8:** Regression-based model performance for all window sizes (1, 2, 3, 4, 5, 10, 20 min) and regression based models (LR, DTR, RFR, XGBR, Base) including labels from writing tasks. Bold marked results of models outperformed the baseline model of the corresponding window size in terms of MSE and had a $R^2 > 0$.

| Window size | Classifier | Flow writing | | Flow writing [normalized] | | Cognitive load writing | | Cognitive load writing [normalized] | |
|---|---|---|---|---|---|---|---|---|---|
| | | MSE | $R^2$ | MSE | $R^2$ | MSE | $R^2$ | MSE | $R^2$ |
| 1 | LR | 1.750 | −4.903 | 0.245 | −15.985 | 0.481 | −3.419 | 0.067 | −0.268 |
| 1 | DTR | **0.916** | **0.071** | 0.082 | −0.405 | 0.693 | −3.062 | 0.077 | −0.408 |
| 1 | RFR | 0.913 | −0.026 | 0.093 | −0.827 | 0.441 | −2.716 | 0.058 | −0.138 |
| 1 | XGBR | 0.859 | −0.071 | 0.104 | −1.465 | 0.559 | −3.017 | 0.060 | −0.169 |
| 1 | Base | 1.049 | −0.006 | 0.091 | −0.566 | 0.495 | −3.584 | 0.067 | −0.251 |
| 2 | LR | 1.132 | −0.176 | 0.105 | −1.204 | 0.552 | −3.274 | 0.091 | −0.603 |
| 2 | DTR | 1.159 | 0.076 | 0.089 | −0.330 | 0.505 | −3.130 | 0.081 | −0.641 |
| 2 | RFR | 1.066 | 0.177 | 0.092 | −0.535 | 0.559 | −3.315 | 0.067 | −0.257 |
| 2 | XGBR | **0.985** | **0.229** | 0.120 | −1.880 | 0.765 | −5.368 | 0.062 | −0.147 |
| 2 | Base | 1.060 | −0.009 | 0.094 | −0.606 | 0.508 | −3.480 | 0.064 | −0.185 |
| 3 | LR | 1.163 | −0.349 | 0.096 | −0.112 | 0.509 | −3.296 | 0.068 | −0.253 |
| 3 | DTR | 1.136 | −0.016 | **0.084** | **0.104** | 0.637 | −3.572 | 0.065 | −0.184 |
| 3 | RFR | 1.136 | −0.706 | 0.087 | −0.334 | 0.577 | −3.880 | 0.064 | −0.252 |
| 3 | XGBR | 1.271 | −1.235 | 0.107 | −0.131 | 0.627 | −3.858 | 0.063 | −0.267 |
| 3 | Base | 1.082 | −0.068 | 0.092 | −0.117 | 0.493 | −3.450 | 0.062 | −0.155 |
| 4 | LR | 1.100 | −0.071 | **0.087** | **0.318** | 0.567 | −3.558 | 0.073 | −0.335 |
| 4 | DTR | 2.067 | −1.661 | 0.095 | −0.051 | 0.554 | −3.809 | 0.068 | −0.254 |
| 4 | RFR | 1.160 | −0.186 | 0.093 | −0.002 | 0.558 | −3.842 | 0.066 | −0.253 |
| 4 | XGBR | 1.502 | −0.498 | 0.112 | −0.184 | 0.585 | −4.039 | 0.077 | −0.404 |
| 4 | Base | 1.062 | −0.072 | 0.090 | −0.136 | 0.486 | −3.390 | 0.062 | −0.151 |
| 5 | LR | 1.176 | −0.499 | 0.097 | −0.537 | 0.525 | −3.473 | 0.066 | −0.260 |
| 5 | DTR | 1.224 | −0.292 | 0.123 | −2.467 | 0.522 | −3.465 | 0.063 | −0.179 |
| 5 | RFR | 1.144 | −0.112 | 0.107 | −1.037 | 0.571 | −3.919 | 0.065 | −0.221 |
| 5 | XGBR | 1.270 | −0.623 | 0.105 | −0.746 | 0.691 | −4.048 | 0.072 | −0.301 |
| 5 | Base | 1.049 | −0.042 | 0.090 | −0.199 | 0.481 | −3.418 | 0.061 | −0.151 |
| 10 | LR | 1.414 | −2.010 | 0.159 | −6.632 | 0.547 | −5.032 | 0.074 | −0.428 |
| 10 | DTR | 1.420 | −0.379 | 0.102 | −0.534 | 0.597 | −5.925 | 0.098 | −0.894 |
| 10 | RFR | 1.095 | −0.132 | 0.100 | −0.711 | 0.483 | −4.860 | 0.067 | −0.394 |
| 10 | XGBR | 1.155 | −0.577 | 0.115 | −0.693 | 0.631 | −6.488 | 0.084 | −0.699 |
| 10 | Base | 1.006 | −0.042 | 0.089 | −0.226 | 0.451 | −4.745 | 0.059 | −0.186 |
| 20 | LR | 3.685 | −19.345 | 0.395 | −30.206 | 0.806 | −5.511 | 0.195 | −2.434 |
| 20 | DTR | 1.154 | −0.004 | 0.122 | −1.197 | 0.436 | −3.955 | 0.067 | −0.261 |
| 20 | RFR | 1.056 | −0.288 | 0.092 | −0.409 | 0.475 | −4.295 | 0.070 | −0.385 |
| 20 | XGBR | 1.158 | −0.227 | 0.127 | −0.398 | 0.460 | −3.709 | 0.097 | −0.724 |
| 20 | Base | 1.006 | −0.042 | 0.089 | −0.226 | 0.451 | −4.745 | 0.059 | −0.186 |

model in terms of MSE (MSE = 0.985). In addition, Decision Tree based regressors also outperformed the baseline regression for a window size of 1 min (MSE = 0.916, $R^2$ = 0.071). For normalized flow labels, Linear Regression based models and a window size of 4 min performed best for $R^2$ ($R^2$ = 0.318) and also outperformed the baseline regression model in terms of MSE (MSE = 0.087). For cognitive load, none of the regression-based models achieved a lower MSE than a baseline model while maintaining a positive $R^2$ value.

Overall, the results of the classification- and regression-based cognitive state models (flow and cognitive load) for writing tasks demonstrate that cognitive state models based on eye-tracking data collected in the field exhibit robustness, accuracy and generalizability across participants. Moreover, these results highlight that cognitive state models are generalizable across participants but not across tasks.

# 5 Discussion

Our results show that developing generalizable, accurate, and robust cognitive state models based on field eye-tracking data is a challenging but feasible task. We were unable to develop an eye-based cognitive state ML model that generalizes across tasks and participants, as the developed models using labels from all tasks did not significantly outperform the baseline classifier in terms of F1-Score and AUC, or the baseline regression in terms of MSE and $R^2$. Eye-tracking data is known to be highly task dependent, which may explain the poor performance of the eye-based cognitive state models when considering labels from all tasks. However, we were successful in developing eye-based cognitive state models when considering only labels collected during writing tasks, as they outperformed the baseline classifier in terms of F1-Score and AUC, and the baseline regression in terms of MSE and $R^2$. Thus, for a writing task, we were able to develop accurate models that achieve generalizability in terms of working across different participants. Furthermore, this suggests that it is possible to develop cognitive state models using eye-tracking data collected in the field.

To further advance the development of eye-based cognitive state models, we systematically examined the eye tracking and log data collected, the labels collected, the model performance, and the interviews with participants. On this basis, we derived six major lessons learned (LLs) from our field study that can help other researchers in conducting field data collection studies with eye tracking for the purpose of building cognitive state models. The first three LLs relate to the data collection method and system, and the subsequent LLs relate to model development using eye-tracking data.

Many existing ESM studies have focused on collecting data on mobile devices.[44] This has typically required shifting attention from the task to another platform. Therefore, we implemented esmLoop to integrate ESM data collection directly into the operating system of the PC. With our approach and the esmLoop system we were able to successfully collect a large amount of data during our study. Using esmLoop, all participants were able to collect large amounts of valid eye tracking (>150 h), interaction, and labeling data (293 labels of cognitive load and flow) on their own, without the active supervision of an experimenter. According to our interviews, participants found esmLoop user-friendly and liked its integration into the operating system. In addition, they reported that resuming tasks after completing surveys was quick because it was well integrated into their work environment. However, even though the system was tightly integrated and worked well on different PCs during testing, participants reported some problems with the software running on their computer during the field studies, which ultimately affected data collection. In the case of participant P12, the eye-tracking driver software stopped working and recording data during the study, resulting in the participant's removal from the experiment. In addition, several users reported that the software crashed due to random disconnections of the eye tracker, causing some frustration for the user. This may be because the students' computer hardware is not up to date, or because the external eye-tracking hardware is not well integrated with the hardware and operating system, or is not designed to record for several hours. It also shows that tight integration of hardware and software with the system is important for robust and high quality data collection in our context. Hardware integration may change as computing devices with a built-in eye tracker, such as the Apple Vision Pro, come to market and continuously rely on eye tracking for interactions, improving integration within the system. These findings underscore our first lesson:

**LL1:** *To make data collection more robust and ensure high data quality, label and eye-tracking data collection should be tightly integrated with the system the user is using to perform their task(s).*

Eye tracking and self-reported cognitive state data is known to be sensitive data, as it is considered health data by law and reveals a lot of information about the user. 72 % of the users who participated in our experiment had no privacy concerns when using esmLoop during the study. They

appreciated the ability to schedule the recording sessions by starting and stopping them as needed, and that the data was not automatically shared for analysis. During the data collection phase, users could see that data collection was in progress from the icon in the task bar. They found that this visual cue gave them a sense of control over their own data and transparency in the process. In addition, when using esmLoop, users were curious about their contribution and found the reports on the number of labels and the amount of data collected to be interesting features of esmLoop, which increased their motivation to continue collecting data. Finally, providing participants with information about the goal of the experiment, the privacy check, the experimental design approved by the ethics board, etc. during the initial workshop also helped to increase their confidence and motivation to actively participate in the experiment. On this basis, we articulate the following lesson learned:

**LL2:** *For the collection of eye tracking and self-reported cognitive state data, which are sensitive data and require high effort, the system should follow a transparent data collection policy to support user privacy and increase motivation to provide data and labels.*

The feeling of being tracked is known to be one of the most common discussions about eye-tracking studies and the potential to bias user behavior. Despite initial concerns about the presence of the eye tracker and the potential distraction of being recorded, as well as the light from the infrared light sources of the eye tracker, users reported that they became accustomed to the presence of the eye tracker and that it did not affect their behavior. In fact, some participants reported the positive effect of being more focused on their task due to the eye tracker and felt that their behavior, and specifically the importance of experiencing the flow state, was important to this research goal. We summarize this finding in the following lessons learned:

**LL3:** *Study participants will become accustomed to the presence of eye trackers and to being recorded.*

Similar to affect detection based on biosignal data collected in the field,[57] developing a generalizable cognitive state ML model using eye-tracking data collected in the field is challenging. To limit the complexity, we decided to approach the model development as a binary classification and regression problem rather than a multi-class problem. While the label distribution analysis shows that cognitive load and flow are experienced differently, this is also supported by the interview data. Furthermore, it is still challenging to define a generally valid threshold between high and low cognitive load and flow or no flow state for training the classification models. Although the visual behavior of the participants was individual, the classification models had a hard time identifying patterns for different cognitive states across tasks that would lead to successful classification. Even our regression models, which did not consider any cognitive state label distribution, did not perform significantly better than the baseline model when considering data from all participants and tasks. Therefore, we would suggest the approach of building within-subject models that, on the other hand, require a large number of labels to cover all manifestations of cognitive states for different tasks from one individual. In our study, we did not collect enough labels per participant to follow this suggestion. Therefore, this suggestion needs to be evaluated in future studies. Based on the above findings, we articulate the following lessons learned:

**LL4:** *Developing an eye-based cognitive state ML model that is generalizable across users is challenging because eye movements are very individual and cognitive states are perceived very differently by individuals.*

The results of our study showed that we were able to collect labels for a wider range of tasks, but for most tasks we collected only a few labels. The model development results suggest that the development of a cognitive load and flow model generalizable across tasks and participants using real-world data was not significantly successful. None of the developed models significantly outperformed the baseline model when considering labels of all tasks. Only a Decision Tree classifier for cognitive load slightly outperformed the baseline model in terms of F1-Score and AUC but exhibited a low discriminative power. Moreover, one regression models based on decision tree regressors for flow slightly outperformed the baseline model, but the $R^2$ value was very low, indicating that little of the variation in cognitive state is predictable from the eye-based features. Only when considering labels collected during a writing task were we able to develop models generalizable across participants that outperformed the baseline models for flow and cognitive load. This is also reflected in the fact that eye movements are highly dependent on task and environment, as shown in Yarbus.[58] To achieve generalizability of eye-based cognitive state models across tasks, not only a large dataset covering a variety of tasks and environments is needed, but also a variety of cognitive state labels need to be collected for each of the tasks.

**LL5:** *Developing an eye-based cognitive state ML model that is generalizable across tasks and environments is challenging because eye movements are highly task and environment dependent. Therefore, it is important to collect enough labels per task and cover a variety of environments*.

At first glance, the performance of the developed cognitive state models looks quite good especially when following a binary classification-based approach. As shown in the Tables 5 and 7, some classifiers achieved an F1-Score above 70 %. However, without a balanced dataset in terms of labels and a comparison with a baseline classifier, an evaluation based on the F1-Score may overestimate the performance of the classifiers. For example, if a baseline classifier for flow with a window size of 1 min also achieves an F1-Score of 72.80 % due to unbalanced data (see Table 5), a Random Forest classifier that achieves an F1-Score of 72.80 % and AUC of 0.5000 is performing just equally and is not learning from the data. One approach to overcome the risk of overestimating performance is to oversample the minority class using SMOTE to create a balanced dataset. This approach proved to be helpful as seen in the Table 5, since the Random Forest classifier for flow and a window size of 1 min did not achieve a higher F1-Score than the baseline classifier after using SMOTE. Furthermore, checking the confusion matrices of the LOOCV showed that the Random Forest classifier did not provide any discriminative power to distinguish between the two classes of flow and no flow, as the classifier always predicted the no flow class for all iterations of the CV, which is also supported by the AUC value of 0.5. In summary, we draw the following lessons learned:

**LL6:** *For the development of eye-based cognitive state models following a binary classification approach, a differentiated evaluation of classifier performance is important*.

### 5.1 Limitations & future work

The findings of this study are limited to the context of thesis work by students. For the present study, we recruited 11 students as subjects who were working on their thesis projects. The students were recruited from different study programs and they worked on different tasks depending on the different stages of their thesis projects. Furthermore, we only collected data during a limited time frame of 5 days, during which we probably could not cover all tasks of that specific thesis work stage. The interview results show that users adjusted their break schedule based on when survey questions might appear. Therefore, there is a need to more dynamically adjust the frequency and timing of survey questions so that users do not change their behavior. In addition, we only examined the cognitive states of cognitive load and flow. There are many more cognitive states, such as situation awareness, comprehension, distraction, certainty, or fatigue, that are also worth investigating using ESM and eye-tracking data to build supervised ML models.

The approach we used to develop the eye-based cognitive state models also has several limitations. Because this was a field study, we focused on a rather broad set of tasks, neglecting the collection of a balanced data set for each task and participant. Despite collecting data over 5 days, we were not able to collect enough labels per participant (approximately 26 labels per participant) to develop individual models for each participant. In the future, we propose to collect more labels per participant and task to investigate the development of within-subject cognitive state models. In addition, we simplified the classification problem to a binary problem and a regression problem, but did not take a multi-class approach. Also, the models developed based on labels collected only during the writing tasks are less representative. They only consider data from a label subset of all participants, since not all participants reported cognitive state labels for the writing tasks.

## 6 Conclusions

In the future, advances in sensor technology and machine learning will make it possible to monitor our cognitive state like our physical activity. Eye-tracking technology is a promising off- and on-body sensor technology that can provide access to cognitive user states. We contribute to this vision by investigating the development of cognitive state models using eye-tracking data collected in the field. In this paper, we present and apply an experience sampling system called esmLoop to record eye-tracking data and collect cognitive state labels from 11 students working on their thesis project in the field. We develop cognitive load and flow models using supervised machine learning algorithms for classification and regression and the collected field data. We also evaluate these models for accuracy, robustness, and generalizability across tasks and users. Our results demonstrate that developing cognitive state models that are generalizable across participants and tasks is challenging, and we have not been successful in developing such models. However, the results of task-specific cognitive state models highlight that it is possible to develop cognitive state models that are generalizable across participants using eye-tracking data collected in the field. Finally, we articulate

six lessons learned during data collection and model development to enable the development of cognitive state models that are generalizable across participants and tasks in the future.

**Acknowledgments:** Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – GRK2739/1 – Project Nr. 447089431 – Research Training Group: KD$^2$School – Designing Adaptive Systems for Economic Decisions. The authors thank Pascal Roller for his support in this research project.

**Research ethics:** The local Institutional Review Board approved the study.

**Author contributions:** The authors have accepted responsibility for the entire content of this manuscript and approved its submission.

**Competing interests:** The authors state no conflict of interest.

**Research funding:** Deutsche Forschungsgemeinschaft, GRK2739/1 – Project Nr. 447089431

**Data availability:** The aggregated data can be downloaded from the cited resource provided in the paper.

## References

1. Davern, M.; Shaft, T.; Te'eni, D. Cognition Matters: Enduring Questions in Cognitive IS Research. *J. Assoc. Inf. Syst.* **2012**, *13* (4), 273—314.
2. Neisser, U. *Cognitive Psychology*, Classic ed.; Psychology Press: New York, 2014.
3. Kosch, T.; Karolus, J.; Zagermann, J.; Reiterer, H.; Schmidt, A.; Woźniak, P. W. A Survey on Measuring Cognitive Workload in Human-Computer Interaction. *ACM Comput. Surv.* **2023**, *55* (13s), 1—39.
4. Wilson, M. L.; Midha, S.; Maior, H. A.; Cox, A. L.; Chuang, L. L.; Urquhart, L. D. SIG: Moving from Brain-Computer Interfaces to Personal Cognitive Informatics. In *Conference on Human Factors in Computing Systems — Proceedings*, 2022; pp. 4—7.
5. Schultz, T.; Maedche, A. Biosignals Meet Adaptive Systems. *SN Appl. Sci.* **2023**, *5* (9), 234.
6. Hutt, S.; Krasich, K.; Brockmole, J. R.; D'Mello, S. K. Breaking Out of the Lab: Mitigating Mind Wandering with Gaze-Based Attention-Aware Technology in Classrooms. In *Conference on Human Factors in Computing Systems — Proceedings*; ACM: New York, NY, USA, 2021; pp. 1—14.
7. Langner, M.; Toreini, P.; Maedche, A. Leveraging Eye Tracking Technology for a Situation-Aware Writing Assistant. In *2023 Symposium on Eye Tracking Research and Applications, ETRA '23*; ACM: New York, NY, USA, 2023; pp. 1—2.
8. Seitz, J.; Krisam, C.; Benke, I. A State of the Art Overview on Biosignal-Based User-Adaptive Video Conferencing Systems. In *Wirtschaftsinformatik 2023 Proceedings*, Vol. *27*, 2023.
9. Langner, M.; Toreini, P.; Maedche, A. EyeMeet: A Joint Attention Support System for Remote Meetings. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts, CHI EA '22*; ACM: New York, NY, USA, 2022; pp. 1—7.
10. Toreini, P.; Langner, M.; Maedche, A.; Morana, S.; Vogel, T. Designing Attentive Information Dashboards. *J. Assoc. Inf. Syst.* **2022**, *22* (2), 521—552.
11. Appel, T.; Scharinger, C.; Gerjets, P.; Kasneci, E. Cross-Subject Workload Classification Using Pupil-Related Measures. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research and Applications — ETRA*, Vol. *18*, 2018; pp. 1—8.
12. Halverson, T.; Estepp, J.; Christensen, J.; Monnin, J. Classifying Workload with Eye Movements in a Complex Task. In *Proceedings of the Human Factors and Ergonomics Society*, 2012; pp. 168—172.
13. Steichen, B.; Carenini, G.; Conati, C. User-Adaptive Information Visualization: Using Eye Gaze Data to Infer Visualization Tasks and User Cognitive Abilities. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces, IUI '13*; ACM: New York, NY, USA, 2013; pp. 317—328.
14. Wang, W.; Li, Z.; Wang, Y.; Chen, F. Indexing Cognitive Workload Based on Pupillary Response under Luminance and Emotional Changes. In *International Conference on Intelligent User Interfaces, Proceedings IUI*, 2013; pp. 247—256.
15. Hoppe, S.; Loetscher, T.; Morey, S. A.; Bulling, A. Eye Movements during Everyday Behavior Predict Personality Traits. *Front. Hum. Neurosci.* **2018**, *12* (April), 1—8.
16. Shiffman, S.; Stone, A. A.; Hufford, M. R. Ecological Momentary Assessment. *Annu. Rev. Clin. Psychol.* **2008**, *4* (1), 1—32.
17. van Berkel, N.; Goncalves, J.; Lovén, L.; Ferreira, D.; Hosio, S.; Kostakos, V. Effect of Experience Sampling Schedules on Response Rate and Recall Accuracy of Objective Self-Reports. *Int. J. Hum. Comput. Stud.* **2019**, *125*, 118—128.
18. Larson, R.; Csikszentmihalyi, M. The Experience Sampling Method. In *Flow and the Foundations of Positive Psychology*; Springer: Dordrecht, 2014; pp. 21—34.
19. Beal, D. J. ESM 2.0: State of the Art and Future Potential of Experience Sampling Methods in Organizational Research. *Annu. Rev. Organ. Psychol. Organ. Behav.* **2015**, *2*, 383—407.
20. Majaranta, P.; Bulling, A. Eye Tracking and Eye-Based Human—Computer Interaction. In *Advances in Physiological Computing*; Fairclough, S.; Gilleade, K., Eds.; Springer: London, 2014; pp. 39—65.
21. Just, M. A.; Carpenter, P. A. A Theory of Reading: From Eye Fixations to Comprehension. *Psychol. Rev.* **1980**, *87* (4), 329—354.
22. Duchowski, A. T. *Eye Tracking Methodology*; Springer Nature: Cham, 2017.
23. Barral, O.; Lallé, S.; Guz, G.; Iranpour, A.; Conati, C. Eye-Tracking to Predict User Cognitive Abilities and Performance for User-Adaptive Narrative Visualizations. In *Proceedings of the 2020 International Conference on Multimodal Interaction, ICMI '20*; Association for Computing Machinery: New York, NY, USA, 2020; pp. 163—173.
24. Berkovsky, S.; Taib, R.; Koprinska, I.; Wang, E.; Zeng, Y.; Li, J.; Kleitman, S. Detecting Personality Traits Using Eye-Tracking Data. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19*; Association for Computing Machinery: New York, NY, USA, 2019; pp. 1—12.
25. Bozkir, E.; Geisler, D.; Kasneci, E. Person Independent, Privacy Preserving, and Real Time Assessment of Cognitive Load Using Eye Tracking in a Virtual Reality Setup. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, 2019; pp. 1834—1837.

26. Taib, R.; Berkovsky, S.; Koprinska, I.; Wang, E.; Zeng, Y.; Li, J. Personality Sensing: Detection of Personality Traits Using Physiological Responses to Image and Video Stimuli. *ACM Trans. Interact. Intell. Syst.* **2020**, *10* (3), 1—32.
27. Conati, C.; Lallé, S.; Rahman, M. A.; Toker, D. Comparing and Combining Interaction Data and Eye-Tracking Data for the Real-Time Prediction of User Cognitive Abilities in Visualization Tasks. *ACM Trans. Interact. Intell. Syst.* **2020**, *10* (2), 1—41.
28. Raptis, G. E.; Katsini, C.; Belk, M.; Fidas, C.; Samaras, G.; Avouris, N. Using Eye Gaze Data and Visual Activities to Infer Human Cognitive Styles: Method and Feasibility Studies. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization, UMAP '17*; Association for Computing Machinery: New York, NY, USA, 2017; pp. 164—173.
29. Ekman, P. An Argument for Basic Emotions. *Cognit. Emot.* **1992**, *6* (3—4), 169—200.
30. Alhargan, A.; Cooke, N.; Binjammaz, T. Multimodal Affect Recognition in an Interactive Gaming Environment Using Eye Tracking and Speech Signals. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction, ICMI '17*; Association for Computing Machinery: New York, NY, USA, 2017; pp. 479—486.
31. Seitz, J.; Maedche, A. *Biosignal-Based Recognition of Cognitive Load: A Systematic Review of Public Datasets and Classifiers*; Springer: Cham, Vol. 43, 2022; pp. 35—52.
32. Hart, S. G.; Staveland, L. E. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Adv. Psychol.* **1988**, *52* (C), 139—183.
33. Kahneman, D.; Beatty, J. Pupil Diameter and Load on Memory. *Science* **1966**, *154* (3756), 1583—1585.
34. Duchowski, A. T.; Krejtz, K.; Krejtz, I.; Biele, C.; Niedzielska, A.; Kiefer, P.; Raubal, M.; Giannopoulos, I. The Index of Pupillary Activity: Measuring Cognitive Load Vis-À-Vis Task Difficulty with Pupil Oscillation. In *Proc. of CHI*, 2018; pp. 1—13.
35. Krejtz, K.; Duchowski, A. T.; Niedzielska, A.; Biele, C.; Krejtz, I. Eye Tracking Cognitive Load Using Pupil Diameter and Microsaccades with Fixed Gaze. *PLoS One* **2018**, *13* (9), 1—23.
36. Abbad-Andaloussi, A.; Sorg, T.; Weber, B. Estimating Developers' Cognitive Load at a Fine-Grained Level Using Eye-Tracking Measures. In *Proceedings of the 30th IEEE/ACM International Conference on Program Comprehension*; ACM: New York, NY, USA, 2022; pp. 111—121.
37. Appel, T.; Gerjets, P.; Hoffman, S.; Moeller, K.; Ninaus, M.; Scharinger, C.; Sevcenko, N.; Wortha, F.; Kasneci, E. Cross-Task and Cross-Participant Classification of Cognitive Load in an Emergency Simulation Game. In *IEEE Transactions on Affective Computing*, 2021; p. 1.
38. Csikszentmihalyi, M. Flow: The Psychology of Optimal Experience: Steps Toward Enhancing the Quality of Life. *Des. Issues* **1991**, *8* (1), 314.
39. Nakamura, J.; Csikszentmihalyi, M. Flow Theory and Research. In *Oxford Handbook of Positive Psychology*, 2009; pp. 195—206.
40. Knierim, M. T.; Bartholomeyczik, K.; Nieken, P.; Weinhardt, C. Could We Predict Flow from Ear-EEG? In *2022 10th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos, ACIIW 2022*, 2022; pp. 1—6.
41. Rissler, R.; Nadj, M.; Li, M. X.; Loewe, N.; Knierim, M. T.; Maedche, A. To Be or Not to Be in Flow at Work: Physiological Classification of Flow Using Machine Learning. *IEEE Trans. Affect. Comput.* **2020**, *14* (1), 463—474.
42. Trull, T. J.; Ebner-Priemer, U. W. Using Experience Sampling Methods/Ecological Momentary Assessment (ESM/EMA) in Clinical Assessment and Clinical Research: Introduction to the Special Section. *Psychol. Assess.* **2009**, *21* (4), 457—462.
43. Kapoor, A.; Horvitz, E. Experience Sampling for Building Predictive User Models: A Comparative Study. In *Conference on Human Factors in Computing Systems — Proceedings*, 2008; pp. 657—666.
44. Van Berkel, N.; Ferreira, D.; Kostakos, V. The Experience Sampling Method on Mobile Devices. *ACM Comput. Surv.* **2017**, *50* (6), 1—40.
45. Karapanos, E. Technology-Assisted Reconstruction: A New Alternative to the Experience Sampling Method. *Behav. Inf. Technol.* **2020**, *39* (7), 722—740.
46. Schmidt, P.; Reiss, A.; Dürichen, R.; Laerhoven, K. V. Wearable-Based Affect Recognition—A Review. *Sensors* **2019**, *19* (19), 4079.
47. Compton, R. J.; Gearinger, D.; Wild, H. The Wandering Mind Oscillates: EEG Alpha Power Is Enhanced During Moments of Mind-Wandering. *Cognit. Affect Behav. Neurosci.* **2019**, *19* (5), 1184—1191.
48. Hutt, S.; Mills, C.; White, S.; Donnelly, P. J.; D'Mello, S. K. The Eyes Have It: Gaze-Based Detection of Mind Wandering During Learning with an Intelligent Tutoring System. In *Proceedings of the 9th International Conference on Educational Data Mining, EDM 2016*, 2016; pp. 86—93.
49. van Berkel, N.; Goncalves, J.; Hosio, S.; Kostakos, V. Gamification of Mobile Experience Sampling Improves Data Quality and Quantity. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, Vol. *1* (3), 2017; pp. 1—21.
50. Rheinberg, F.; Vollmeyer, R.; Engeser, S. *FKS-Flow-Kurzskala*; ZPID (Leibniz Institute for Psychology Information), Testarchiv: Trier, 2019.
51. Almalki, K.; Alharbi, O.; Al-Ahmadi, W.; Aljohani, M. Anti-procrastination Online Tool for Graduate Students Based on the Pomodoro Technique. In *Learning and Collaboration Technologies. Human and Technology Ecosystems*; Zaphiris, P.; Ioannou, A., Eds.; Springer International Publishing: Cham, 2020; pp. 133—144.
52. Dalmaijer, E. S.; Mathôt, S.; Stigchel, S. V. D. PyGaze: An Open-Source, Cross-Platform Toolbox for Minimal-Effort Programming of Eye-Tracking Experiments Edwin. *Behav. Res. Methods* **2014**, *46*, 1—16.
53. Langner, M.; Toreini, P.; Maedche, A. *Cognitive State Detection with Eye Tracking in the Field: An Experience Sampling Study and its Lessons Learned - Dataset & Analysis Script*; Karlsruhe Institute of Technology: Karlsruhe, 2024.
54. Bethge, D.; Chuang, L.; Grosse-Puppendahl, T. Analyzing Transferability of Happiness Detection via Gaze Tracking in Multimedia Applications. In *ACM Symposium on Eye Tracking Research and Applications, ETRA '20 Adjunct*; Association for Computing Machinery: New York, NY, USA, 2020.
55. Kaczorowska, M.; Plechawska-Wójcik, M.; Tokovarov, M. Interpretable Machine Learning Models for Three-Way Classification of Cognitive Workload Levels for Eye-Tracking Features. *Brain Sci.* **2021**, *11* (2), 1—22.
56. Hutt, S.; Stewart, A. E.; Gregg, J.; Mattingly, S.; D'Mello, S. K. Feasibility of Longitudinal Eye-Gaze Tracking in the Workplace. In

*Proceedings of the ACM on Human-Computer Interaction*, Vol. *6* (ETRA), 2022; pp. 1—21.
57. D'Mello, S. K.; Booth, B. M. Affect Detection from Wearables in the "Real" Wild: Fact, Fantasy, or Somewhere In Between? *IEEE Intell. Syst.* **2023**, *38* (1), 76—84.
58. Yarbus, A. L. *Eye Movements and Vision*; Plenum Press: New York, 1967.

## Bionotes

**Moritz Langner**
Karlsruhe Institute of Technology, Karlsruhe, Germany
**moritz.langner@kit.edu**
**https://orcid.org/0000-0001-7860-7118**

Moritz Langner is a PhD candidate at the Karlsruhe Institute of Technology (KIT). His research interest is focused on Human-Computer Interaction and the design of user-adaptive systems using eye tracking technology to support students in managing their cognition.

**Peyman Toreini**
Karlsruhe Institute of Technology, Karlsruhe, Germany
**peyman.toreini@kit.edu**
**https://orcid.org/0000-0002-2468-1715**

Peyman Toreini is a post-doctoral researcher at the Karlsruhe Institute of Technology (KIT). His research focuses on human-computer interaction by designing and developing the next generation of user interfaces for working environments. In particular, he is interested in the novel user experiences that are enabled by eye trackers on desktop stations, smart glasses, augmented and mixed reality environments.

**Alexander Maedche**
Karlsruhe Institute of Technology, Karlsruhe, Germany
**alexander.maedche@kit.edu**
**https://orcid.org/0000-0001-6546-4816**

Prof. Dr. Alexander Maedche is a full professor of Information Systems (IS) and Human-Computer Interaction (HCI) at the Karlsruhe Institute of Technology (KIT) in Germany. He is heading the human-centered systems lab (h-lab) with a focus on designing human-centered systems for better work and life. He publishes in leading outlets in IS and HCI, like Management Information Systems Quarterly, Information Systems Research, IEEE Transactions on Affective Computing, CHI, and CSCW. He actively promotes transfer to practice and is co-founder of the non-profit organizations Usability and User Experience in Germany (UIG) e.V. and Die Wirtschaftsinformatik e.V.