Loughborough University

This item was submitted to Loughborough's Research Repository by the author.
Items in Figshare are protected by copyright, with all rights reserved, unless otherwise indicated.

Cross-task and cross-participant classification of cognitive load in an emergency simulation game

PLEASE CITE THE PUBLISHED VERSION

https://doi.org/10.1109/TAFFC.2021.3098237

PUBLISHER

Institute of Electrical and Electronics Engineers

VERSION

AM (Accepted Manuscript)

PUBLISHER STATEMENT

Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

LICENCE

All Rights Reserved

REPOSITORY RECORD

Appel, Tobias, Peter Gerjets, Stefan Hoffman, Korbinian Moeller, Manuel Ninaus, Christian Scharinger, Natalia Sevcenko, Franz Wortha, and Enkelejda Kasneci. 2021. “Cross-task and Cross-participant Classification of Cognitive Load in an Emergency Simulation Game”. Loughborough University.

Cross-task and Cross-participant Classification of Cognitive Load in an Emergency Simulation Game

Tobias Appel, *Student Member, IEEE,* Peter Gerjets, Stefan Hoffmann, Korbinian Moeller, Manuel Ninaus, Christian Scharinger, Natalia Sevcenko, Franz Wortha, Enkelejda Kasneci

Abstract—Assessment of cognitive load is a major step towards adaptive interfaces. However, non-invasive assessment is rather subject as well as task specific and generalizes poorly, mainly due to methodological limitations. Additionally, it heavily relies on performance data like game scores or test results. In this study we present an eye-tracking approach that circumvents these shortcomings and allows for generalizing well across participants and tasks. First, we established classifiers for predicting cognitive load individually for a typical working memory task (n-back), which we then applied to an emergency simulation game by considering the k most similar ones and weighting their predictions. Standardization steps helped achieving high levels of cross-task and cross-participant classification accuracy between 63.78% and 67.25% for the distinction between easy and hard levels of the emergency simulation game. These very promising results may pave the way for novel adaptive computer-human interaction across domains and particularly for gaming and learning environments.

Index Terms—Eye Tracking, Physiology, Intelligent Systems, Cognitive Model, Physiological Measures, Psychology, Adaptive and Intelligent Educational Systems.

1 Introduction

In many digital environments designed for learning, working or even for entertainment purposes there is a close link between users' current cognitive load and their affective experiences. An important impact of users' cognitive load on their affective states has been demonstrated repeatedly (e.g. [1], [2], [3]). For instance, in a learning context, it is highly relevant for users' subjective experiences that the instructions provided by a learning environment are neither over-straining (and thereby frustrating) for learners nor under-challenging (and thus boring) due to a lack of workload imposed [4]. However, imposing an optimal level of cognitive load onto learners that keeps them engaged and satisfied as well as in their zone of proximal development [5] might be a highly learner-specific issue that strongly depends on individual learners' prerequisites in terms of their prior knowledge and abilities. Similar to the zone of proximal development, the Yerkes-Dodson law suggests to manage arousal for optimal performance. Since cognitive load can be interpreted as a form of arousal - as illustrated by its impact on pupil diameter - managing it would be beneficial in many situations. For calibrating learning experiences with regard to their cognitive demands and affective connotations technical support systems might be helpful. Such a system could optimize the resulting learning outcomes by performing real-time adjustments of the cognitive-load level imposed by a learning environment.

Systems able to detect and properly react to a user's cognitive load in order to calibrate their affective and cognitive experiences would of course also offer considerable benefits for numerous other applications beyond learning environments - ranging from the workplace to the digital playground. For instance, in potentially stressful digital working environments (such as systems for surgical assistance, engine control or emergency management), in which errors might have serious and life-threatening consequences, individuals might also experience strong and fluctuating affective reactions related to their current level of cognitive (over-)load. Monitoring cognitive load in these contexts and providing respective feedback and support to users might not only help to avoid errors related to high cognitive load but also to improve the overall affective experience. For instance, (truck) drivers or other workers controlling complex engines may be prompted to take breaks or can be provided with individualized training when detected to be over-strained in specific situations (e.g. [6] or [7]). Other examples might be conceivable in the medical domain where surgeons might be relieved when necessary, or in aviation scenarios where pilots may be provided with support from their copilots or from assistance systems depending on their

Tobias Appel was with LEAD Graduate School and Research Network, Tübingen, Germany. He is now with the Department of Human-Computer Interaction, University of Tübingen, Tübingen , Germany.
Peter Gerjets is with the Leibniz-Institut für Wissensmedien, Tübingen, Germany.
Stefan Hoffman is with Promotion Software GmbH, Tübingen, Germany.
Korbinian Moeller was with the Leibniz-Institut für Wissensmedien, Tübingen. He is now with the Centre for Mathematical Cognition, University of Loughborough, UK.
Manuel Ninaus is with the University of Innsbruck, Innsbruck, Austria.
Christian Scharinger is with the Leibniz-Institut für Wissensmedien, Tübingen, Germany.
Natalia Sevcenko is with the Leibniz-Institut für Wissensmedien, Tübingen, Germany and with Daimler AG, Stuttgart, Germany.
Franz Wortha was with LEAD Graduate School and Research Network, Tübingen, Germany. He is now with the Leibniz-Institut für Wissensmedien, Tübingen, Germany.
Enkelejda Kasneci is with the Department of Human-Computer Interaction, University of Tübingen, Tübingen, Germany.
Manuscript received TBA;

cognitive-load levels [8].

Beyond scenarios related to learning or working, gaming also seems to be a prime area for applying adaptive procedures based on cognitive load measurement in order to optimize affective user experiences. For instance, in cases where an obstacle in a game is too difficult for a player to overcome, frustration may set in and the gaming experience may suffer. Contrarily, when a game is too easy in relation to users current abilities, gaming may also not be experienced as enjoyable. Both cases can be circumvented by adapting the degree of difficulty based on the player's current level of cognitive load. Thus, with accurate estimations of cognitive load during gaming, automatic adaptations might be enabled that prevent negative affective states (such as boredom, frustration, and stress) and enhance positive affects (such as engagement, joy, and satisfaction). A prime example of a desirable affective state in gaming and many other scenarios of human-computer interaction is flow [9]. Flow is considered a positive affective state of optimal experience [10] that creates pleasure by balancing the challenge of the task at hand and the available capabilities of the user. Measuring a person's cognitive load online might help to adapt the levels of difficulty to a degree where it still constitutes enjoyable challenge but does not over-strain the user.

Usually, cognitive load is measured based on self-reports, such as the NASA task-load index (TLX) [11], or by obtaining performance metrics. These traditional approaches, however, have some drawbacks that render them impractical for systems aiming at real-time adaptations. In particular, filling out a questionnaire about the level of cognitive load currently experienced might strongly interfere with task performance and immersion and is therefore unsuitable for most applications. Moreover, questionnaire data are rather subjective and may be influenced by many factors with the current level of cognitive load being only one of them [12]. Thus, for real-time adaptations to cognitive load levels, a less obtrusive method would be required that does not interrupt the current task like questionnaires but offers a reliable, objective and continuous indirect online estimation of user's cognitive load.

Performance metrics such as test scores or task completion times are indirect and thus less interfering compared to questionnaires. However, they are usually only available at specific points in time and can not be measured continuously as would be required for real-time adaptations to cognitive load levels. For instance, in the case of digital learning environments, one would aim at measuring cognitive load levels during the learning process for adapting difficulty levels of learning materials and not only after a learning task is completed (e.g. by means of a test). Thus, the required continuous and unobtrusive cognitive-load monitoring can usually neither be provided by performance metrics nor by questionnaire data [13].

An alternative approach for assessing cognitive load is based on physiological measures. Cognitive load causes physiological reactions that can be measured by sensors [13], [14], [15]. The most reliable indicators are changes that occur in the brain, but measuring these changes is intrusive, hard to set up, and not feasible in broad real-world settings. In this context, methods like electroencephalography (EEG) or near-infrared spectroscopy (NIRS) require very specialized hardware and expertise to operate. A thorough overview on EEG measures, including their advantages and drawbacks - such as the number of required trials per experiment - that illustrate why these measures are rather unsuitable in the context of most real-life adaptive systems is provided by [16]. Less intrusive sensors include heart rate monitors and devices for measuring skin conductance, which however seem to lack accuracy and/or validity for measuring cognitive load [17]. Finally, eye tracking measures such as eye-fixation features offer a good alternative to the aforementioned physiological signals. They do not require physical contact with participants, can be obtained in real-time, and have been comprehensively demonstrated to be associated with cognitive load [18]. When obtained by means of webcams, eye-tracking measures have the potential to become available to a broad audience across various application domains. Moreover, with the increasing integration of eye-tracking technology in VR, AR, and smart glasses [19], this physiological signal can also be measured in high quality in a variety of applications in the future [20].

One of the major limitations of physiological indicators such as eye-tracking data for measuring cognitive load consists in the difficulty to generalize measures across tasks and across participants [21], rendering cross-task and cross-participant predictions or real-time assessments virtually impossible. This drawback, however, is not limited to eye-tracking data or physiological measures in general but applies to many algorithms for real-time workload assessment as Heard et al. conclude in their meta-review [22]. Systems designed for real-time assessments usually need to be adapted to individual participants and/or specific tasks in order to yield reliable predictions. This usually requires data collection for lengthy calibration procedures for each participant rendering these systems time consuming and inconvenient for users. Even with individual calibration, generalizations to different tasks or applications are usually poor, resulting in a necessity for repeated calibrations for different tasks and/or applications. Currently, there is no satisfying general purpose classifier for cognitive load available.

Many researchers have worked on the problem of either cross-task or cross-participant estimations of cognitive load (see below for a more detailed discussion), but with limited success so far. Usually, although intra-participant results are good generalization results are limited or do not even exceed chance level (e.g. [23], [24], [25]).

In this article, we present a novel and intuitive approach how to remedy these methodological shortcomings. We show how a machine learning approach might be used for cognitive load detection based on eye-tracking data to allow for successful generalization across participants and tasks. We employ a schema of weighted votes that combines participant-specific classifiers into a composite classifier with a broader scope offering generalization ability across participants and tasks. Our method might thus be able to pave the ground for out-of-the-box solutions for adaptive human-computer interaction based on a reliable assessment and classification of users' cognitive load independent of the user and task at hand. As a result, users' affective experiences during human-computer interaction in contexts

such as learning, working, or gaming might strongly benefit in terms of avoiding frustration, boredom, or stress and in terms of enhancing engagement, joy, and satisfaction. In line with this assumption we show that our cognitive-load classification is significantly correlated with negative emotions such as stress and frustration.

2 Related Work

2.1 Adaptations based on cognitive load estimation

Cognitive load estimation usually is performed in a task and participant-specific way and has also been demonstrated in this context to allow for useful workload adaptations in learning environments or vehicle control tasks. Yuksel and colleagues created a brain-computer interface that adapted the difficulty of a musical learning task [26]. They measured cognitive load using fNIRS to decide when to increase difficulty. Their approach managed to significantly increase learning gains during piano lessons compared to a control group. However, the classifiers they used were participant-specific and were trained using a long training period consisting of 30 songs per participant.

Moreover, an aviation simulation was used by Wilson and Russel to provide real-time adaptive feedback [8]. They used a combination of EEG, respiration, and heart rate, but also eye-fixation behavior to realize adaptations during an uninhabited air vehicle task. Participant-specific artificial neural networks were trained to detect high cognitive load and adapt the task by slowing down simulated time when cognitive load was too high. In contrast to our approach, real-time adaptation was successfully realized only with participant- and task-specific classifiers.

Furthermore, Kelleher et al. developed a method that does not rely on EEG data, but rather on users' behavioural performance [27]. Their approach was able to distinguish between a hard puzzle and an easier one based on users' performance on the previous puzzles with an accuracy of 71% to 79%. A wide array of features derived from performance, user input, and user ratings was used to train random forests and predictions were made based on the last three puzzles the user was attempting to solve. While the results are promising, their method is still specific to their task and individual participants.

2.2 Cross-participant and cross-task approaches

While cognitive load estimation usually is performed in a task and participant-specific way, there are several studies that successfully implemented either cross-task or cross-participant approaches (but not both). In contrast to the method that we present in this article, most of them rely, at least partly, on EEG.

A very detailed assessment of mental workload is provided by Popovic et al. [28]. They classified different kinds of load (i.e., speech, fine motor, gross motor, auditory, visual and cognitive) using EEG and ECG. Their cross-participant classifier achieved 72.5% accuracy for cognitive load in a leave-one-participant-out cross-validation.

Another interesting approach was presented by Ke and colleagues [29]. They generalize from individual regression models to more general ones by applying a feature selection algorithm to EEG data recorded from a working memory task and a complex simulated multi-attribute task designed to evaluate operator performance and workload (see [30]). In a first step, they used two thirds of their data to systematically eliminate features with low cross-task correlations and then evaluated their feature set on the remaining validation set. They found a significant increase in performance of their regression model. Again, these results show cross-task- but not cross-participant-generalization. This makes them applicable in some situations, but still not as general as a many applications would demand.

Finally, in previous work our group successfully developed a machine learning approach for cross-participant classification of cognitive load [31] using eye-tracking data. For a working memory task we achieved an accuracy of 76.8% for offline classification and 70.4% for real-time online classification. A reworked version of this approach was used in an emergency simulation and showed promising results under noisy conditions similar to actual applications [32]. This updated version worked across different versions of the simulation, showing potential for a cross-participant and cross-task solution. The article presented here expands on this work by refining the set of used eye-tracking features and adding a further weighing step for cross-task application that makes use of the accuracy scores. The previous two articles were limited to one task or variations of one task, but in this work we perform actual cross-task and cross-participant classification working towards a truly general algorithm.

3 Experimental setup

We collected eye-tracking data from two different tasks: (1) an N-back task (a standardized working memory task inducing a controlled level of cognitive load), and (2) a computer simulation, that represented a real-life application. Pursuing cross-task classification we aimed to use data from the first to estimate cognitive load in the latter, thereby strictly separating the training task and validation task. This separation was crucial in order to keep the approach as general as possible. We further ensured that there is no overlap between the two groups of participants to guarantee that it also is strictly cross-participant. Both are crucial aspects of this work.

3.1 N-back task

The N-back task [33] is commonly used to induce cognitive load and to measure working memory capacity. Participants are presented with a randomly generated sequence of letters and have to press one of two buttons to indicate whether the currently presented letter is the same as N letters before. N modulates the difficulty of the task, because a larger N means that more letters have to be held in memory and compared to the actual one presented. With regard to working-memory demands, the N-back task requires to keep a string of N letters active in memory, compare the first letter of the string to the current trial, decide on the correct button, and update the memorized string by deleting the first letter of the string and adding the letter of the current trial to it. 0-back can be used as a control condition where

participants have to compare the current stimulus with a constant that was presented at the very beginning. Letters are randomly chosen from the set $L = \{C, F, H, S\}$ and are presented for 0.5 seconds, followed by a black screen shown for 1.5 seconds. A schematic overview is provided in Figure 1 and descriptive statistics can be found in 2.

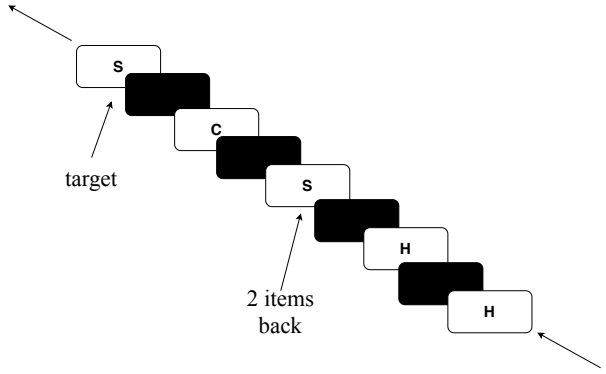

Fig. 1. Overview of the N-back task as illustrated for $N = 2$

Participants first received instructions for the task and had to perform a short training until they achieved an accuracy of 60%. They then completed two critical blocks, each comprising three difficulty levels: 0-back, 1-back, and 2-back. Each level of a block consisted of 154 trials; the order of levels in a block was randomized.

We used the "N" of the N-back task as an experimental manipulation of cognitive load (within participants design) and focused on the difference between 0-back and 2-back conditions.

3.1.1 Participants

28 students (*mean age* = 24.71, *SD* = 4.12, 14 females) from the University of Tübingen were recruited for the N-back task. Data of one participant was discarded due to problems with the eye-tracking recordings resulting in too little usable data.

The experiment was approved by the local ethics committee and all participants gave written informed consent at the beginning of the experiment. Participants received monetary compensation at the end of the experiment. All were right-handed and German native speakers.

3.1.2 Apparatus

We used a RED250 eye tracker from SensoMotoric Instruments (SMI) in combination with the SMI Experiment Center software (version 2.7.13) for the recording of eye movements and pupil-related features. Calibration was performed with SMI's built-in 9-point calibration. All eye-tracking data were recorded at 250Hz in a laboratory setting with illumination held constant in individual sessions.

During the task a chin-rest was used to ensure stable head position and constant viewing distance. Stimuli were presented on a 22-inch monitor with a resolution of 1,680 $\times$ 1,050 px using Arial font with a size of 25. All letters were presented in gray on a black background.

3.2 *Emergency* simulation task

The *Emergency* simulation task was based on the commercially available simulation *Emergency* by Promotion Software GmbH [34]. Participants had to coordinate emergency personnel consisting of firefighters, paramedics, and ambulances involved in responding to different scenarios (e.g., a car crash or burning buildings). The simulation can be used as a training tool for emergency management tasks as well as for entertainment purposes and thus covers aspects of digital environments for learning, working, and gaming.

The simulation started with a tutorial that introduced participants to the handling of the simulation. After the tutorial was completed successfully, three scenarios were presented: A car crash, burning buildings, and a train crash. Each scenario had three levels of increasing difficulty (easy, medium, and hard version of the scenario). Scenarios had to be completed in the same order of easy to difficult levels and scenarios (i.e., from car crash to train crash) by all participants. Scenarios and difficulty levels differed in the number of sub-tasks to be completed as well as the available numbers of emergency personnel and their composition. These manipulations were calibrated by Promotion Software GmbH for the purposes of this study in order to optimally manipulate the levels of cognitive load imposed onto participants. In order for a scenario to be completed successfully, all sub-tasks must be performed, meaning all trapped victims had to be freed, every injured person had to be treated and transported to a hospital, and every fire had to be extinguished. Descriptive statistics about the scenarios' parameter, completion rates, and subjective ratings can be found in Table 1.

The fixed order of scenarios and difficulty was chosen deliberately for this task. While a randomized order would be ideal to avoid confounds, it is hard to implement in a task that involves learning and skill acquisition. Participants who complete easy parts of the simulation first gain proficiency fast and scenarios that may have been difficult for them in the beginning become more and more easy. On the other hand, when participants are confronted with very difficult scenarios first, they may be overwhelmed which hinders or even prevents learning. This expertise change over time and its dependency on the order of task presentation necessitated a fixed order. Gerjets et al. also recommend a fixed order from simple to complex for learning tasks for this exact reasons [4]. Moreover, an ascending order of difficulty is in line with the way learning materials and games are commonly structured, making it a better showcase for the application of our method. A further important aspect is that we aim to use data from the N-back task to estimate cognitive load in the *Emergency* simulation so that the training data are not affected by a confounding of difficulty level and time.

In the simulation certain sub-tasks could only be performed by specific emergency personnel and with varying degrees of efficiency. Therefore, especially in scenarios that did involve fire, planning activities were essential for successful completion of a mission. For instance, as fire can spread to nearby buildings and also hurt emergency personnel, prioritisation of which fires to put out first was crucial. Putting out fires could be done by firetrucks but

Fig. 2. This image is taken from the third scenario "train crash" and shows a typical situation in *Emergency*. Paramedics and ambulances can be seen at the bottom of the screen, as well as the firefighters' trucks.

also by firefighters alone, however, with firefighters being considerably slower at performing the task. Moreover, firefighters might be required for cutting trapped victims free. In general, more difficult levels involved more emergency personnel units to be coordinated and more sub-tasks, posing higher demands on planning, prioritisation, monitoring, and information updating.

After each level, participants were asked to indicate their subjective cognitive load and also their affective experiences based on a modified version of the NASA-TLX questionnaire [11]. The questionnaire contained scales for positive and negative emotions, mental and temporal demand, effort, frustration, and stress, as well as a measure of seriousness that were each rated on a scale of 0 to 100. Participants' responses were used to evaluate the validity of our approach and the relation of our envisioned cognitive-load classification to affective experiences.

For this task, we used the difficulty levels within each scenario as manipulation of cognitive load (within participants design) and focused on the difference between the easy and hard version of each scenario.

3.2.1 Participants

The Emergency simulation was completed by 47 participants (*mean age* = 24.6, *SD* = 6.3, 33 females). There was no overlap between the participants of the *Emergency* simulation and the participants of the N-back task. Seven participants had to be excluded due to problems with eye-tracking recordings. Another 2 were excluded because they reported that they did not take the experiment seriously. Finally, 2 participants had a very high number of missing data and consequently did not provide enough usable data for all scenarios, rendering their data partly unusable. The data of the remaining 36 participants were included in the further analyses.

We deliberately included participants with noisy data or poor tracking ratios (i.e., time spans with invalid data caused by the pupil not being detected reliably). This renders the data more realistic with closer resemblance to data one would expect in an online-scenario of a real-world application.

The experiment was approved by the local ethics committee and all participants gave written informed consent at the beginning of the experiment. All participants were right-handed, German native speakers and received monetary compensation at the end of the experiment.

3.2.2 Apparatus

The eye-tracking setup was the same as for the N-back task, featuring a RED250 eye tracker from SensoMotoric Instruments (SMI) in combination with the SMI Experiment Center software (3.7.60). Calibration was performed with SMI's built-in 9-point calibration and the recording frequency was set to 250Hz. Data recording was performed in a laboratory setting with illumination held constant in individual sessions.

For the *Emergency* simulation a laptop with a 16-inch screen driven at 1920 x 1080 px resolution was used. This task did not involve a chin-rest as to closer mimic a real-world learning or gaming situation.

4 Features used for Classification

Eye-fixation behavior is strongly influenced by presented stimuli as their structure and appearance guide the users

Scenario	difficulty	units	sub-tasks	fire	time	completed	pupil	cog. demand	temp. demand	effort
scenario 1	easy	6	11	no	300s	83.33%	4.05	35.85	32.43	33.29
scenario 1	hard	12	20	no	300s	83.33%	4.15	42.14	40.29	38.14
scenario 2	easy	9	18	yes	450s	97.22%	4.15	39.00	23.14	32.57
scenario 2	hard	15	26	yes	450s	36.11%	4.30	51.14	56.57	52.71
scenario 3	easy	10	24	yes	600s	97.22%	4.13	43.43	31.57	39.43
scenario 3	hard	16	44	yes	600s	33.33%	4.25	59.00	64.57	61.86

TABLE 1
Descriptive statistics of the three scenarios of the Emergency simulation game.

N-back	pupil diameter [mm]	accuracy	reaction time [ms]
level 0	5.59	88%	462
level 1	5.72	86%	506
level 2	6.14	79%	632

TABLE 2
Descriptive statistics of the three scenarios of the Emergency simulation game.

attention (e.g., Rayner, for a review [35], [36]). Therefore, our approach relies on eye-related features that were chosen because they are either independent of the stimulus structure or only marginally depended on it. More specifically, we did not rely on saccades, areas of interest, or the coordinates of fixations.

The feature extraction process for a chosen share of data always followed the same procedure. First we extracted 7 features which will be described later in this section and then normalized them using a participant-specific baseline to allow for cross-participant comparisons. Baseline in this context refers to the features of a specific part of the data. For the N-back task this baseline was taken from the instruction phase, while for the Emergency simulation we used the tutorial phase as baseline.

Normalization was performed at the participant level and involved subtracting the baseline from the segment's features and dividing by it. As a consequence, all features that we used reflected relative changes from the individual participant's baseline.

All eye event detection used SMI's built-in methods. For fixations this is a dispersion-based algorithm with a maximum dispersion of $2-3^\circ$ (depending on the distance between screen and user) and a minimum fixation duration of 80ms. Blinks are defined via the gaze and pupil signal. Gaze coordinates of $(0,0)$ or the pupil being zero or outside a dynamic computed validity range is interpreted as a blink. Blinks of less than 70ms are discarded. SMI's default algorithm interprets anything that is between between two fixations or a blink and a fixation as a saccade. Even though we did not use saccades for our approach, we included their detection for completeness sake.

4.1 Pupil-related features

Pupil diameter has been used to measure cognitive load for several decades. An increase in cognitive load leads to decreased parasympathetic activity in the peripheral nervous system, which in turn leads to an increase in pupil diameter [37]. This effect was observed consistently within a task, between tasks, and between individuals [38]. Various studies have successfully replicated this relationship within a wide range of settings, including short-term memory, language processing, reasoning, perception, as well as sustained and selective attention [18]. Pupil diameter has also successfully been used to detect cognitive load in a variety of scenarios, including driving [39], during low visual load tasks [40], route planning with maps [41], and simultaneous interpreting [42]. Furthermore, it was successfully used to differentiate expertise closely related to cognitive load [43]

We applied preprocessing steps to improve data quality of the pupil signal. First, we removed periods that were marked as blinks, as well as the 100ms right before and after a blink. During these phases, the pupil could not be detected reliably and as a consequence measurements of pupil diameter would suffer from reduced accuracy. We furthermore removed implausible pupil values (e.g., values of 0mm or less, as well as values greater than 10mm). Finally, we linearly interpolated small gaps of less than 50ms (12 data points at sampling rate 250Hz) and applied a median filter to reduce noise.

We selected the median of the pupil diameter as the main pupil feature, as it is more robust to outliers than the mean, in particular for short sampling periods. Moreover, we utilized the maximum pupil diameter as a feature to capture spikes in the pupil signal. We expected to see an increase in both median and maximum pupil diameter with increasing cognitive load.

Moreover, we employed the Index of Cognitive Activity (ICA) as proposed by Marshall [44], [45]. It uses wavelet decomposition to detrend a pupil signal and reduce it to only short-term fluctuations where rapid pupil spikes that exceed a threshold can be detected. These spikes are supposedly caused by cognitive activity. For this part of the analysis, we did not apply any interpolation or filter preprocessing as it may remove the small-scale fluctuations needed for the ICA. A higher degree of cognitive load was supposed to result in an increased ICA.

As an additional, more exploratory feature, we included the standard deviation of the pupil diameter. According to the ICA, cognitive load can cause fluctuations and rapid spikes in pupil diameter. Based on this assumption, we expected a higher standard deviation of pupil diameter for higher cognitive load.

4.2 Blinks

Cognitive load influences the frequency and duration of blinks [46], [47]. Thus, increasing task difficulty was expected to increase the frequency of blinks, while increasing visual demands should lower the amount of blinks [48]. We used blink frequency as a feature and expected it to increase with cognitive load in both the N-back task and the *Emergency* simulation.

4.3 Fixations

Fixations describe a stable gaze on the same location usually lasting between 200 ms and 350 ms [35]. Frequency of fixations is influenced by several factors. Time pressure induced by high task demands tends to increase the number of fixations while reducing their duration [49]. We expected to observe the same pattern for higher levels of cognitive load in our study. Consequently, we used the number of fixations per second as a feature.

4.4 Microsaccades

Microsaccades are small involuntary eye movements that may occur during a fixation and are associated with cognitive load. Studies reported an increase in microsaccade frequency in visually demanding tasks [50], whereas non-visual tasks (e.g, auditory tasks or mental arithmetic) seemed to reduce their frequency [51], [52], [53].

We used the method suggested by Krejtz and colleagues [53] to detect microsaccades, which relies on thresholds to find small ballistic sequences in an otherwise fixed gaze. Instead of focusing on amplitude or velocity we use microsaccade frequency as a feature, as the former two would require a higher sampling rate than 250Hz to be reliable. Because both tasks involved visual presentation, we expected an increase in microsaccade frequency with rising cognitive load.

5 Cognitive Load Detection Method

The core of our approach was strongly inspired by Appel et al. [31], [32]. The fundamental idea was to train participant-specific classifiers for low and high cognitive load based on data from a N-back task and use their weighted predictions on the Emergency simulation. Participants that were similar during baseline periods are weighted stronger as we expect their physiology to change under cognitive load in a similar way.

5.1 Within-task and within-participant classification

Participant-specific classifiers were trained on N-back data. As the N-back is a standard working-memory updating task that is recorded under laboratory conditions, we expected it to reflect characteristic physiological changes caused by cognitive load and to allow for generalization from this task to the *Emergency* simulation as described in Section 5.2.

To train a classifier that can differentiate between high and low cognitive load, we needed data from periods of high cognitive load and periods of low cognitive load during the training phase. We used the N-back task as foundation for single-participant classifiers and considered the 0-back condition to reflect low cognitive load and the 2-back condition to represent high cognitive load. 25 non-overlapping samples with a length of 4s each were randomly selected from both conditions, yielding 50 samples per participants that were used for training the individual classifier. We rejected samples with more than 50% missing values in the pupil signal and resampled to ensure that each sample contained enough information to be useful for training. These numbers represented a balanced compromise between sample size and sample length.

For each of the samples we extracted the features described in Section 4. All features were then z-transformed using individual means and SDs for standardization. This scaling improved inter-participant comparability considerably and should thus help applying classifiers across participants.

Finally, we trained a forest of 1000 extremely randomized trees (Extra-Trees) [54] per participant to distinguish between high and low cognitive load based on that individual participants' samples. Extra-Trees had the advantage of providing not just a decision into classes, but also class probabilities between 0 and 1. This enabled us to form a continuous scale instead of a dichotomous decision, adding further information. The output was a number between 0, when the classifier was absolutely certain that a sample was collected under low cognitive load, and 1 in case of high cognitive load. In addition, Extra-Trees tended to not overfit as fast as other classification methods allowing more features in conjunction with less samples. Moreover, Extra trees seemed appropriate for the goal of real-time classification of cognitive-load levels as they can be trained and evaluated fast. The use of 1000 trees was empirically determined. More trees make the classifier more robust and more accurate, but require longer time for training. On the one hand, adding trees beyond 1000 did not increase accuracy, neither within-participant nor cross-participant, but on the other hand reducing the number did not improve training or evaluation time in a meaningful way. In case computation time is an issue, less trees may be chosen as to ensure acceptable execution times.

We used the Extra-Tree implementation provided by the Python toolbox scikit-learn [55].

5.2 Cross-participant and cross-task approach

In a next step, we combined the single-participant classifiers trained with data from the N-back task to form a composite classifier that can be applied across participants and tasks. The fundamental idea of our approach was to apply the classifiers trained on N-back data to the Emergency simulation, but to weigh their contribution to the final prediction according to how similar their baselines were. In this way, participants from the N-back task that were more similar regarding their physiological features and behavioral parameters to a participant from the *Emergency* task were given higher weights in the final prediction. Adding this weighing substantially increased the accuracy of the combined classifier.

To verify cross-task capability, sample data from the *Emergency* task was needed. Therefore, we randomly sampled 25 segments of length 4s from the easy and hard version of each scenario of *Emergency*, resulting in 50 samples per scenario and participant. Again, these numbers represented a compromise between sample length and sample number. From these samples we extracted the features in the same way as we did with the N-back data including the normalization using the baseline and z-transformation. Segments extracted from the easy version of a scenario represented low cognitive load, while segments from the hard version represented high cognitive load.

For baseline comparison, features were normalized across all participants to have a mean of 0 and standard

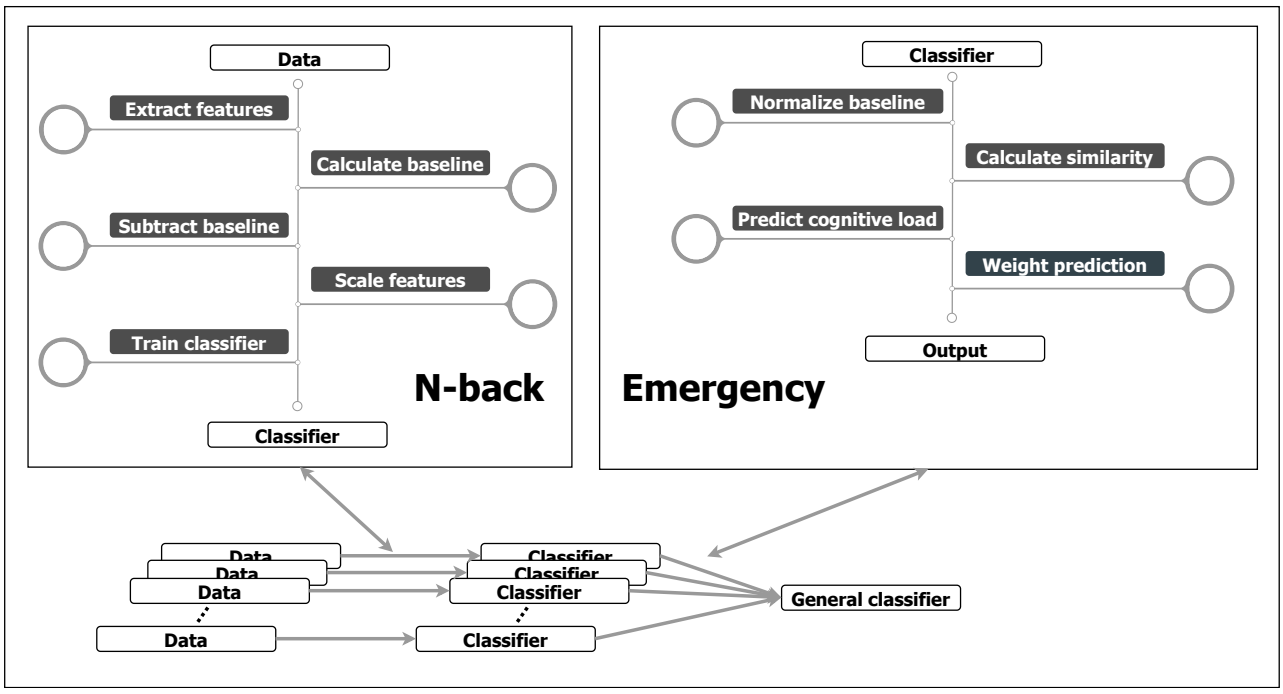

Fig. 3. Overview of our method, showing how we train individual classifiers and apply them to new data.

deviation of 1 as to not inflate the importance of features that are on a larger scale. There were, for instance, a lot fewer blinks within one second than there were fixations and the pupil diameter in millimeters was a lot larger compared to the number of microsaccades per second.

The procedure can be described as follows (see 3 for an overview of all variables): Let p be a participant of *Emergency* whose cognitive load we want to classify, $base_p(i)$ the ith baseline feature, $sample_p$ a sample of p characterized by a set of features, and P the set of participants of the N-back task. Every $q \in P$ has a classifier c_q that predicts a value between 0 and 1 for $sample_p$. We combined these predictions according to the following equations:

$$sim(p,q) = \frac{1}{\sum_i acc_{c_q} w_{c_q(i)} |base_p(i) - base_q(i)|}$$

$$pred(sample_p) = \frac{\sum_{q \in P_n} sim(p,q) pred_{c_q}(sample_p)}{\sum_{q \in P_n} sim(p,q)}$$

$sim(p,q)$ refers to the baseline similarity between participants p and q, $w_{c_q}(i)$ to the ith normalized feature weight of classifier c_q, acc_{c_q} to the cross-validated accuracy that classifier c_q achieved on participant q, and $pred_q$ to the prediction of classifier c_q. This means that we drew a prediction for cognitive load from each classifier c_q and weighted these predictions according to how similar the baselines of participants p and q are. Additionally, we factored the feature weights of classifier c_q into the similarity, giving a higher weight to more important features. Feature weights can be obtained by examining the trees of a random forest and how the addition of a specific feature reduces impurity (see [56] for more details). We further factor in how well the classifier performed on its specific participant by multiplying by its accuracy score. This way, participants with more exemplary data - and consequently a good participant-specific accuracy - are weighted higher.

Dividing by the sum of all similarities normalized these similarities and ensured that the prediction's final result was within the interval of $[0, 1]$. P_n refers to a subset of P, that is restricted to the n participants with the highest similarity. The choice of a smaller n can help to reduce computational costs in case there are a lot of classifiers available from the N-back task. We employed $n = 5$ to highlight that it does not take a lot of participants to get accurate results.

variable	description
p	participant of *Emergency*
q	participant of N-back
c_q	classifier trained on data of participant q
acc_{c_q}	accuracy of c_q
w_{c_q}	feature weights of c_q
$sample_p$	a sample of participant p
$pred_{c_q}(x)$	prediction of c_q for a sample x
$base_p$	baseline features of participant p
$base_p$	baseline features of participant q
$sim(p,q)$	baseline similarity between participants p and q

TABLE 3
Description of all variables in formulae and pseudo-code.

Figure 3 shows a schematic overview of our method and Algorithm 1 presents pseudo-code of our cross-participant and cross-task classification. Both serve to illustrate our approach.

6 Results

6.1 Within-task

As a frame of reference, we did not only analyze results for cross-task and cross-participant classification but also

Algorithm 1 Pseudocode outlining our method for cognitive load detection

$P \leftarrow$ set of all participants in N-back task
$d_{p,i} \leftarrow$ data of participant p taken from the ith scenario of Emergency
$base_P \leftarrow$ normalized baseline of participant p
$c_P \leftarrow$ N-back-trained classifier of participant p
$n \leftarrow$ number of neighbours to consider
for $q \in P$ **do** ▷ calculate distances between participants' baselines and make predictions
  $acc \leftarrow accuracy(c_q)$
  $w \leftarrow featureweights(q)$
  $w \leftarrow \frac{w}{\sum w}$
  $dist(p,q) \leftarrow \sum acc\, w|base_p - base_q|$
  for $i \in \{1,2,3\}$ **do**
    $prediction_{q,p,i} \leftarrow c_q.predict(d_{p,i})$ ▷ prediction for each sample of $d_{p,i}$
  end for
  $dist \leftarrow \frac{dist}{\sum dist}$ ▷ normalize distances to sum to 1
  $sim \leftarrow \frac{1}{dist}$ ▷ get similarity from the distance
  $P_n \longleftarrow \{y \in P | y\ amongst\ n\ most\ similar\}$ ▷ n participants with highest similarity to p
  for $i \in \{1,2,3\}$ **do**
    for $sample \in d_{p,i}$ **do**
      $out_p[i, sample] \longleftarrow \frac{\sum_{q \in P_n} sim(x,q) pred_{c_q}(sample)}{\sum_{q \in P_n} sim(p,q)}$
    end for
  end for
end for

for within-task and within-participant classification. To this end, we performed the method described in Section 5.1 for within-participant results and the approach described in Section 5.2 for cross-participant within-task results, but limited to participants of one task. All results reported for within-participant classification were obtained based on a 10-fold cross-validation to avoid overfiting a classifier to a specific participant and thereby artificially inflating classification accuracy.

Within-task accuracy for the *Emergency* task is reported for each scenario individually and is based on random samples with a length of 4 seconds that were extracted from the easy and hard version, respectively. Feature extraction was performed in the same way as described in detail for the N-back samples.

In case a participant reported that the easy version of a scenario was experienced to be more difficult than the hard version, that scenario was excluded from the results of that participant. "More difficult" refers to the average rating of cognitive demands, temporal demands, and effort as reported in the NASA TLX. For 10 participants, the first scenario had to be excluded for this reason and for 2 the second scenario. In the easy version of the first scenario, participants had their first real interaction with the simulation and it is therefore likely that they experienced it as more difficult than the hard version, because by that time they already were familiar with the simulation.

Table 4 shows the detailed accuracy scores for the N-back task and emergency task, respectively.

It is notable that the results for the N-back task were slightly better than those obtained for the Emergency task. Partly, this may be because of the experimental setup. The N-back task was recorded with participants using a chin-rest, which helped to improve quality of the eye-tracking data in general and reliability of the pupil measurements in particular. Furthermore, difficulty remained constant over the course of one level, whereas situational difficulty varied during levels of the emergency simulation. The fact that we took random samples from easy and difficult versions of the simulation may thus have led to samples not reflecting the exact same degree of cognitive load, even within one participant. This, in turn, added variance to the features and made the labels "easy" and "difficult" less distinct for *Emergency* than for the N-back.

	within-participant	cross-participant
N-back	79.55%	75.81%
Emergency, Scenario 1	71.91%	71.41%
Emergency, Scenario 2	71.98%	69.34%
Emergency, Scenario 3	68.91%	67.16%

TABLE 4
Detailed accuracy of within-task classification.

Comparing the drop in accuracy caused by the shift from within-participant to cross-participant classification, one can see that the drop was more pronounced for the N-back task. This was likely due to the fact that we had a larger number of participants for the Emergency simulation, meaning that it was more likely to find good matches during the baseline comparison.

6.2 Cross-task

Cross-task and cross-participant results were obtained by applying classifiers trained on N-back data to samples from the *Emergency* simulation following the approach described in Section 5.2. Classification accuracy is summarized in Table 5 and ranged between 63.78% and 69.25%.

As expected, applying N-back classifiers to Emergency data led to a slight drop in classification accuracy as it represented a classification across different participants and tasks. Using classifiers trained on participant and applying

Scenario	accuracy
Scenario 1	69.25%
Scenario 2	63.78%
Scenario 3	64.02%

TABLE 5
Accuracy of classifiers trained with N-back data and applied to data gathered during the emergency simulation considering $n = 5$ neighbours during weighted voting.

it to a different one introduced a certain error as the classifier did not match the participant. The same holds true for the application across tasks. Moreover, as Figure 4 shows, feature weights also differed between the two tasks introducing yet another source of error.

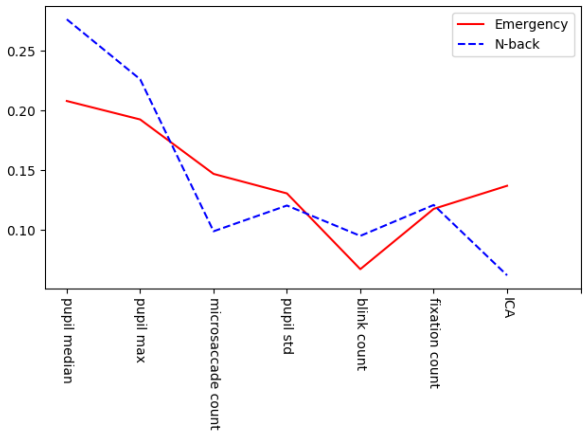

Fig. 4. Average weights for Emergency and N-back tasks

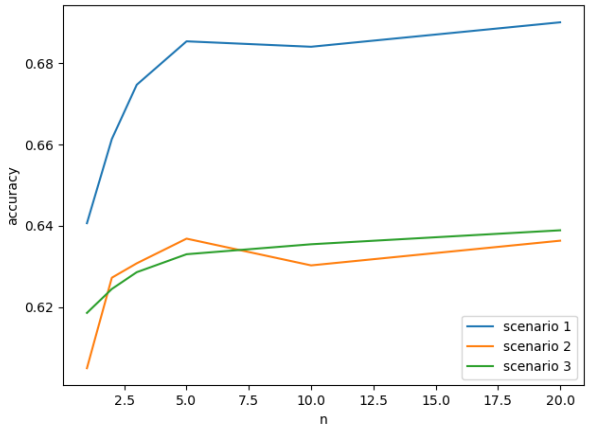

Fig. 5. Classifier performance for cross-task and cross-participant classification dependent on the chosen number of neighbours

The main difference in feature importance was observed for ICA and microsaccades. Both carried considerably more importance in the Emergency simulation than they did for the N-back task. This was in line with results of Fairclough and colleagues who found the ICA to not be significantly sensitive to isolated working memory tasks like the N-back task [57]. A possible explanation for the difference in microsaccade importance may be the different nature of the task. Emergency was a lot more visually demanding and required more widely distributed attention. This fits with findings from Duchowski and colleagues [58] that ambient visual search increases the number of microsaccades. The importance of the remaining features was slightly higher for the N-back task because ICA and microsaccades were less important and the importance of all features sums up to 1.

For scenario 1 classification accuracy decreased from 71.91% from within-participant, within-task application to 69.25% in cross-participant, cross-task classification. A possible explanation for this good performance may be that participants did not yet have experience with the task (i.e., they all started at the same point). This "neutral" conditions with regards to the experience and skills acquired may be similar to the N-back task, thus leading to less loss caused by cross-task application of classifiers.

Scenario 2 showed the most pronounced decrease, from 71.98 to 63.78%. The major error source was the cross-task application, as cross-participant results differed only slightly from the within-participant ones for Emergency. It seems likely that the structure of this specific scenario lead to a feature distribution that differed the most from the N-back task's features, resulting in this decrease in accuracy.

An accuracy loss of less than 5 percentage points – from 68.91% to 64.02% – could be observed for scenario 3. These results were very similar to the second scenario and likely have a similar cause: cross-task application.

Figure 5 depicts the performance of our approach depending on the number of neighbours considered. As expected, there was an increase in accuracy with increasing number of considered neighbours, but after 5 the benefit was negligible only increasing computation time. Even by just choosing the closest match, performance was only a few percentage points lower compared to a larger set of 5 or 10.

To verify our prediction not only on a binary level, we considered participants' questionnaire data and their correlation with our predicted continuous cognitive load. Table 6 shows Pearson correlations between cognitive load predicted by our algorithm and self-report scores. Self-reports were normalized on participant level to account for individual differences of scale. As a frame of reference, we furthermore included correlations between the questionnaire's different sub- scales.

Predictions made by our algorithm showed a significant correlation with self-reports. They correlated at 0.399 with self-reported cognitive demands, at 0.459 with reported temporal demands, and at 0.484 with the effort subjectively experienced by participants. The high correlation with perceived effort is a strong indicator for the validity of our predictions.

7 Discussion

We applied a machine learning approach to the classification of cognitive load based on eye-tracking data and investigated how this approach generalizes across participants and tasks. Our results indicate a robust approach that yields good classification accuracy of 63.78% to 69.25% across participants and tasks. This is above chance level and is comparable to eye-tracking based classification results for

measure	prediction	pos. emotions	neg. emotions	cog. demand	temp. demand	effort	frustration	stress	seriousness
prediction	1.000								
pos. emotions	-0.111	1.000							
neg. emotions	0.186**	0.037***	1.000						
cog. demand	0.399***	-0.134***	0.212***	1.000					
temp. demand	0.459***	-0.175***	0.285***	0.633***	1.000				
effort	0.484***	-0.203***	0.283***	0.701***	0.758***	1.000			
frustration	0.294***	-0.338***	0.443***	0.474***	0.546***	0.607***	1.000		
stress	0.404***	-0.184***	0.309***	0.551***	0.622***	0.657***	0.573***	1.000	
seriousness	0.012	0.175***	-0.112**	-0.009	-0.031	-0.023	-0.058	-0.016	1.000

TABLE 6

*Significant at $p \leq 0.05$, **Significant at $p \leq 0.01$, ***Significant at $p \leq 0.001$

cognitive load in other scenarios. For instance, Hogervost et al. [59] reported roughly 68% accuracy in the distinction between level 0 and level 2 for the N-back task purely based on eye-related features. However, their classification algorithm was trained for each participant individually, within a specific task, and using intervals that were 50s long – all limitations that our approach does not have. Additionally, cognitive load predictions yielded by our method correlate at $r = 0.484$ with participants' self-reported invested effort, which provides an indicator of validity on a second level.

When taking a closer look at our data set, the robustness of our approach seems noteworthy. We considered noisy data in our analyses and, in the case of the *Emergency* game, used the tutorial as an active baseline instead of a neutral fixation cross. Furthermore, *Emergency* is not a well-controlled laboratory task, but a complex emergency simulation game, which requires participants to identify what to do with the right emergency personnel under time constraints in an environment that adaptively reacts to players actions (e.g., spreading fires when not extinguished by fire fighters). As such, the present results seem promising as they indicate the validity of our approach even when applied to a real-world scenario with limited baseline options and complex interactions.

Moreover, due to the dynamic nature of the *Emergency* simulation, cognitive load was not constant over the course of one level. Closer inspection of the predictions generated by our algorithm revealed that participants seem to start each level with rather high predicted load which quickly dropped after a first orientation phase of about 20-30s. Towards the end there was also a clear difference between participants that successfully finished a level and those who did not. When participants realized they will finish on time, predicted cognitive load dropped considerably, whereas it rose when they became aware that they could not finish on time. This uneven distribution of cognitive load adds to the error rates that we report. Therefore, our predictions may actually be even more accurate then what is reported, because we had to rely on the overall task difficulty of a specific level as an indicator of cognitive load instead of a more direct measure (e.g., derived from interaction metrics). Generalizing a difficulty level by labeling all samples from this level as "high cognitive load" even though there were probably periods of lower cognitive load within the same time-frame possibly introduced a kind of artificial error.

Additionally, the nature of our features and method are very versatile. All features we used are either aggregated over the whole length of the segment or are calculated per second. As a result, length of segments can be adjusted at will. Longer segments are less noisy, but shorter segments better capture the cognitive load at a certain point in time. Pre-trained classifiers may be applied independent of segment length, making our approach more flexible. The same holds true for the number of classifiers that are used. When computation time is a constraint, less classifiers may be used for prediction, as n – the number of closest classifies during baseline comparison – can be adjusted at will.

7.1 Limitations

There are, however, also some limitations to our approach. The biggest limitation arises from z-transformation of features on participant level that is required for making data of participants and tasks comparable and on a similar scale. This means that we can only reliably analyze data in hindsight and when there are periods of low and high cognitive load. This scaling also limits the scope with which to interpret cognitive load during different tasks. Only if two tasks share a similar difference in cognitive load across its experimental manipulation can the predictions be considered reliable. A truly objective classifier for cognitive load would need features that are not normalized as to be on the same scale for all participants and circumstances. Our future goal is it to compensate for environmental and individual factors to train a classifier that can objectively estimate the difficulty of a task or activity.

Scaling renders real-time application extremely difficult, too, as it requires a complete dataset to be of use. One may nevertheless use the presented approach in real-time scenarios, but then it has to be considered that workload predictions might not be reliable in the very beginning, but will improve over time as more data becomes available and more variation in cognitive load is observed.

Moreover, one of the reasons why our approach works successfully may also be considered a drawback, namely: baseline comparisons. When the baseline for two tasks is recorded under different conditions problems might arise. For instance, cognitive load may be different in a baseline obtained while looking at a fixation cross as compared to a baseline extracted from completing a tutorial. Using the suggested process of matching participants for cross-participant and cross-task classification, this may lead to a sub-optimal distance metric and consequently an inappropriate weighing of predictions. Ideally, all baselines should evoke the same degree of cognitive load for baseline distances to operate best.

Furthermore, as our analysis of the feature weights showed, a complex simulation game like such as Emergency

does not evoke the exact same physiological responses as a laboratory working- memory task like the N-back task. Our results indicate that in particular the importances of the ICA and microsaccades seems to depend on the task at hand. This hints that – although successful cross-task classification is possible - there may not be a classifier that works optimally for all tasks. In part this may also be a result of the inadequate use of task difficulty as a proxy for cognitive load. Research by Howard et al. suggests that caution has to be exercised comparing difficulty single-task paradigms with multi-tasking ones [60]. A potential solution for this problem could be to have classifiers trained on data of different tasks and also add these tasks to the baseline comparison. This way, classifiers trained on tasks that are similar in nature will be preferred.

Our study relied on visual stimuli and did not include other modalities such as auditory tasks. However, at least for working memory tasks important features seem to react alike and independently of presentation modality. For instance, Kahneman demonstrated extensively and with many different stimuli, tasks, and modalities that working memory load impacts pupil diameter [38]. Recent research yielded similar findings also for microsaccades [51] - at least for auditory and visual stimuli.

7.2 Future Research

Based on this manuscript, many avenues for potential future research are possible. A more elaborate study design that uses the same group of participants across a number of different tasks may be insightful as to what good cognitive load estimators for individual participants entail. It may also help evaluate what features are useful across a wide array of tasks and what features are rather specific to certain types of tasks.

Another way to further this line of research is to find ways to compensate for environmental factors, especially light. Provided that there is an adequate way to do this, we could repeat this experiment without the need to scale features, thereby estimating objective cognitive load instead of cognitive load that is relative to participant and task.

Finally, to make cognitive load estimation possible with systems that are even less intrusive than remote eye trackers and widely available, a webcam-based solution would e ideal. This, however, necessitates webcams that provide good enough resolution to get accurate pupil measurements. Testing with different quality levels of hardware are needed to judge the requirements and feasibility of such an approach. The final goal would be a system that can be employed in everyday life and is able to impact a broad audience allowing for applications in a real-life situation instead of a lab environment.

8 Conclusion

In summary, we evaluated a cross-participant as well as cross-task classification algorithm that yields good accuracy. Combined with the robustness of our method and its non-invasive nature this article - despite its limitations - provides a promising step towards out-of-the-box solutions for adaptive human-computer interaction based on the assessment and classification of users' cognitive load by means of eye-tracking data.

Acknowledgments

This research was funded by the LEAD Graduate School and Research Network [GSC1028], a project of the Excellence Initiative of the German federal and state governments. Tobias Appel and Franz Wortha were doctoral students of the LEAD Graduate School and Research Network.

References

[1] M. S. Hussain, R. A. Calvo, and F. Chen, "Automatic Cognitive Load Detection from Face, Physiology, Task Performance and Fusion During Affective Interference," *Interacting with Computers*, vol. 26, no. 3, pp. 256–268, 06 2013. [Online]. Available: https://doi.org/10.1093/iwc/iwt032

[2] J. M. Rose, F. D. Roberts, and A. M. Rose, "Affective responses to financial data and multimedia: The effects of information load and cognitive load," *International Journal of Accounting Information Systems*, vol. 5, no. 1, pp. 5–24, 2004.

[3] K. Gasper and J. Hackenbracht, "Too busy to feel neutral: Reducing cognitive resources attenuates neutral affective states," *Motivation and Emotion*, vol. 39, no. 3, pp. 458–466, 2015.

[4] P. Gerjets, C. Walter, W. Rosenstiel, and M. Bogdan, "Cognitive state monitoring and the design of adaptive instruction in digital environments: lessons learned from cognitive workload assessment using a passive brain-computer interface approach," *Frontiers in Neuroscience*, 01 2014.

[5] S. Chaiklin, "The zone of proximal development in vygotsky's analysis of learning and instruction," *Vygotsky's educational theory in cultural context*, vol. 1, pp. 39–64, 2003.

[6] L. Zhang, J. Wade, D. Bian, J. Fan, A. Swanson, A. Weitlauf, Z. Warren, and N. Sarkar, "Cognitive load measurement in a virtual reality-based driving system for autism intervention," *IEEE Transactions on Affective Computing*, vol. 8, no. 02, pp. 176–189, apr 2017.

[7] A. Yuce, H. Gao, G. L. Cuendet, and J. Thiran, "Action units and their cross-correlations for prediction of cognitive load during driving," *IEEE Transactions on Affective Computing*, vol. 8, no. 02, pp. 161–175, apr 2017.

[8] G. F. Wilson and C. A. Russell, "Performance enhancement in an uninhabited air vehicle task using psychophysiologically determined adaptive aiding," *Human Factors*, vol. 49, no. 6, pp. 1005–1018, 2007.

[9] M. Csikszentmihalyi, *Flow: The psychology of optimal experience*, vol. 1990.

[10] K. Kiili, A. Lindstedt, and M. Ninaus, "Exploring characteristics of students' emotions, flow and motivation in a math game competition," in *GamiFIN*, 01 2018.

[11] S. G. Hart and L. E. Staveland, "Development of nasa-tlx (task load index): Results of empirical and theoretical research," in *Advances in Psychology*. Elsevier, 1988, vol. 52, pp. 139–183.

[12] R. D. McKendrick and E. Cherry, "A deeper look at the nasa tlx and where it falls short," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 62, no. 1, pp. 44–48, 2018. [Online]. Available: https://doi.org/10.1177/1541931218621010

[13] P. Antonenko, F. Paas, R. Grabner, and T. Van Gog, "Using electroencephalography to measure cognitive load," *Educational Psychology Review*, vol. 22, no. 4, pp. 425–438, 2010.

[14] C. S. Ikehara and M. E. Crosby, "Assessing cognitive load with physiological sensors," in *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, Jan 2005, pp. 295a–295a.

[15] A. Jimenez-Molina, C. Retamal, and H. Lira, "Using psychophysiological sensors to assess mental workload during web browsing," *Sensors*, vol. 18, no. 2, p. 458, 2018.

[16] A. M. Beres, "Time is of the essence: A review of electroencephalography (eeg) and event-related brain potentials (erps) in language research," *Applied psychophysiology and biofeedback*, vol. 42, no. 4, pp. 247–255, 2017.

[17] L. M. Naismith and R. B. Cavalcanti, "Validity of cognitive load measures in simulation-based training: a systematic review," *Academic Medicine*, vol. 90, no. 11, pp. S24–S35, 2015.

[18] J. Beatty, "Task-evoked pupillary responses, processing load, and the structure of processing resources." *Psychological Bulletin*, vol. 91, no. 2, p. 276, 1982.

[19] V. Clay, P. Koenig, and S. Koenig, "Eye tracking in virtual reality," *Journal of Eye Movement Research*, vol. 12, 04 2019.

[20] E. Bozkir, D. Geisler, and E. Kasneci, "Person independent, privacy preserving, and real time assessment of cognitive load using eye tracking in a virtual reality setup," in *The IEEE Conference on Virtual Reality and 3D User Interfaces (VR) Workshops*, mar 2019.

[21] J. L. Lobo, J. D. Ser, F. De Simone, R. Presta, S. Collina, and Z. Moravek, "Cognitive workload classification using eye-tracking and eeg data," in *Proceedings of the International Conference on Human-Computer Interaction in Aerospace*. ACM, 2016, p. 16.

[22] J. Heard, C. E. Harriott, and J. A. Adams, "A survey of workload assessment algorithms," *IEEE Transactions on Human-Machine Systems*, vol. 48, no. 5, pp. 434–451, Oct 2018.

[23] C. L. Baldwin and B. Penaranda, "Adaptive training using an artificial neural network and eeg metrics for within- and cross-task workload classification," *NeuroImage*, vol. 59, no. 1, pp. 48 – 56, 2012, neuroergonomics: The human brain in action and at work. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S105381191100824X

[24] C. Walter, S. Schmidt, W. Rosenstiel, P. Gerjets, and M. Bogdan, "Using cross-task classification for classifying workload levels in complex learning tasks," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, Sep. 2013, pp. 876–881.

[25] M. Spüler, T. Krumpe, C. Walter, C. Scharinger, W. Rosenstiel, and P. Gerjets, "Brain-computer interfaces for educational applications," in *Informational Environments*. Springer, 2017, pp. 177–201.

[26] B. F. Yuksel, K. B. Oleson, L. Harrison, E. M. Peck, D. Afergan, R. Chang, and R. J. Jacob, "Learn piano with bach: An adaptive learning interface that adjusts task difficulty based on brain state," in *Proceedings of the 2016 CHI conference on Human Factors in Computing Systems*. ACM, 2016, pp. 5372–5384.

[27] C. Kelleher and W. Hnin, "Predicting cognitive load in future code puzzles," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI '19. New York, NY, USA: ACM, 2019, pp. 257:1–257:12. [Online]. Available: http://doi.acm.org/10.1145/3290605.3300487

[28] D. Popovic, M. Stikic, T. Rosenthal, D. Klyde, and T. Schnell, "Sensitive, diagnostic and multifaceted mental workload classifier (physioprint)," in *International Conference on Augmented Cognition*, vol. 9183, 08 2015.

[29] Y. Ke, H. Qi, F. He, S. Liu, X. Zhao, P. Zhou, L. Zhang, and D. Ming, "An eeg-based mental workload estimator trained on working memory task can work well under simulated multi-attribute task," *Frontiers in Human Neuroscience*, vol. 8, p. 703, 2014. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnhum.2014.00703

[30] Y. Santiago-Espada, R. R. Myer, K. A. Latorella, and J. R. Comstock Jr, "The multi-attribute task battery ii (matb-ii) software for human performance and workload research: A user's guide," 2011.

[31] T. Appel, C. Scharinger, P. Gerjets, and E. Kasneci, "Cross-subject workload classification using pupil-related measures," in *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, 2018, pp. 1–8.

[32] T. Appel, N. Sevcenko, F. Wortha, K. Tsarava, K. Moeller, M. Ninaus, E. Kasneci, and P. Gerjets, "Predicting cognitive load in an emergency simulation based on behavioral and physiological measures," in *2019 International Conference on Multimodal Interaction*, 2019, pp. 154–163.

[33] W. K. Kirchner, "Age differences in short-term retention of rapidly changing information." *Journal of experimental psychology*, vol. 55, no. 4, p. 352, 1958.

[34] P. S. GmbH. (1999) World of emergency. [Online]. Available: [https://www.world-of-emergency.com/?lang=en

[35] K. Rayner, "Eye movements in reading and information processing: 20 years of research." *Psychological Bulletin*, vol. 124, no. 3, p. 372, 1998.

[36] ——, "Eye movements and attention in reading, scene perception, and visual search," *The quarterly journal of experimental psychology*, vol. 62, no. 8, pp. 1457–1506, 2009.

[37] A. F. Kramer, "Physiological metrics of mental workload: A review of recent progress," *Multiple-task Performance*, pp. 279–328, 1991.

[38] D. Kahneman, *Attention and effort*. Citeseer, 1973, vol. 1063.

[39] O. Palinko, A. L. Kun, A. Shyrokov, and P. Heeman, "Estimating cognitive load using remote eye tracking in a driving simulator," in *Proceedings of the 2010 symposium on eye-tracking research & applications*. ACM, 2010, pp. 141–144.

[40] S. Chen and J. Epps, "Using task-induced pupil diameter and blink rate to infer cognitive load," *Human–Computer Interaction*, vol. 29, no. 4, pp. 390–413, 2014.

[41] P. Kiefer, I. Giannopoulos, A. Duchowski, and M. Raubal, "Measuring cognitive load for map tasks through pupil diameter," in *The Annual International Conference on Geographic Information Science*. Springer, 2016, pp. 323–337.

[42] K. G. Seeber, "Cognitive load in simultaneous interpreting: Measures and methods," *Target. International Journal of Translation Studies*, vol. 25, no. 1, pp. 18–32, 2013.

[43] N. Castner, T. Appel, T. Eder, J. Richter, K. Scheiter, C. Keutel, F. Hüttig, A. Duchowski, and E. Kasneci, "Pupil diameter differentiates expertise in dental radiography visual search," *PLOS ONE*, vol. 15, no. 5, pp. 1–19, may 2020. [Online]. Available: https://doi.org/10.1371/journal.pone.0223941

[44] S. P. Marshall, "Method and apparatus for eye tracking and monitoring pipil dilation to evaluate cognitive activity," Patent US 006 090 051A, 07 18, 2000. [Online]. Available: https://patentimages.storage.googleapis.com/pdfs/9171d27ab488a900c7db/U

[45] ——, "The index of cognitive activity: measuring cognitive workload," in *Proceedings of the IEEE 7th Conference on Human Factors and Power Plants*, 2002, pp. 7–5–7–9.

[46] S. Benedetto, M. Pedrotti, L. Minin, T. Baccino, A. Re, and R. Montanari, "Driver workload and eye blink duration," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 14, no. 3, pp. 199 – 208, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S136984781000094X

[47] K. F. V. Orden, W. Limbert, S. Makeig, and T.-P. Jung, "Eye activity correlates of workload during a visuospatial memory task," *Human Factors*, vol. 43, no. 1, pp. 111–121, 2001, pMID: 11474756. [Online]. Available: http://dx.doi.org/10.1518/001872001775992570

[48] M. Á. Recarte, E. Pérez, Á. Conchillo, and L. M. Nunes, "Mental workload and visual impairment: Differences between pupil, blink, and subjective rating," *The Spanish journal of psychology*, vol. 11, no. 2, pp. 374–385, 2008.

[49] M. De Rivecourt, M. Kuperus, W. J. Post, and L. J. Mulder, "Cardiovascular and eye activity measures as indices for momentary changes in mental effort during simulated flight," *Ergonomics*, vol. 51, no. 9, pp. 1295–1319, 2008.

[50] S. Benedetto, M. Pedrotti, and B. Bridgeman, "Microsaccades and exploratory saccades in a naturalistic environment," *Journal of Eye Movement Research*, vol. 4, no. 2, pp. 1–10, 2011.

[51] X. Gao, H. Yan, and H.-j. Sun, "Modulation of microsaccade rate by task difficulty revealed through between-and within-trial comparisons," *Journal of Vision*, vol. 15, no. 3, pp. 3–3, 2015.

[52] E. Siegenthaler, F. M. Costela, M. B. McCamy, L. L. Di Stasi, J. Otero-Millan, A. Sonderegger, R. Groner, S. Macknik, and S. Martinez-Conde, "Task difficulty in mental arithmetic affects microsaccadic rates and magnitudes," *European Journal of Neuroscience*, vol. 39, no. 2, pp. 287–294, 2014.

[53] K. Krejtz, A. T. Duchowski, A. Niedzielska, C. Biele, and I. Krejtz, "Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze," *PLOS ONE*, vol. 13, no. 9, p. e0203629, 2018.

[54] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 63, no. 1, pp. 3–42, 2006.

[55] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[56] G. Louppe, L. Wehenkel, A. Sutera, and P. Geurts, "Understanding variable importances in forests of randomized trees," *Advances in neural information processing systems*, vol. 26, pp. 431–439, 2013.

[57] S. H. Fairclough, L. J. Moores, K. C. Ewing, and J. Roberts, "Measuring task engagement as an input to physiological computing," in *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, Sep. 2009, pp. 1–9.

[58] A. Duchowski, K. Krejtz, J. Zurawska, and D. House, "Using microsaccades to estimate task difficulty during visual search of layered surfaces," *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–1, 2019.

[59] M. A. Hogervorst, A.-M. Brouwer, and J. B. F. van Erp, "Combining and comparing eeg, peripheral physiology and eye-related measures for the assessment of mental workload," *Front Neurosci*,

vol. 8, p. 322, Oct 2014, 25352774[pmid]. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4196537/

[60] Z. L. Howard, N. J. Evans, R. J. Innes, S. D. Brown, and A. Eidels, "How is multi-tasking different from increased difficulty?" *Psychonomic Bulletin & Review*, 2020.

Tobias Appel received his Bachelor's and Master's degree in Computer Science from the University of Tübingen in 2014 and 2017 respectively. As of 2017 he is pursuing his Ph.D. in Computer Science at the LEAD Graduate School and Research Network and the University of Tübingen. His research focuses on the evaluation of cognitive load based on Eye Tracking and other physiological sensors. In his research he relies on machine learning to realize cross-participant and cross-task solutions.

Peter Gerjets finished his diploma in psychology at the University of Goettingen in 1991. From 1991 to 1995 he was a Research Associate at the University of Göttingen where he received his Ph.D. in 1994. Afterwards he has been working as Assistant Professor at the Saarland University in Saarbrücken where he finished his habilitation in 2002 before taking over his current position at the University of Tübingen. Since 2002 he has been working as principal research scientist at the Knowledge Media Research Center and beside as full professor for research on learning and instruction at the University of Tübingen. He was honoured with the Young Scientist Award of the German Cognitive Science Society in 1999 and served in the editorial boards of the Journal of Educational Psychology, Educational Re-search Review, Computers in Human Behavior, Metacognition and Learning, and Educational Technology, Research, and Development. His current research focuses on multimodal and embodied interaction with digital media as well as on learning from multimedia, hypermedia, and the Web. He is a member of DGPs, APS, and EARLI and served as coordinator of the EARLI Special Interest Group 6: Instructional Design.

Stefan Hoffmann Stefan Hoffmann is a developer of video games for more than 30 years now, with a strong focus on Serious Games and research projects with Serious Games and/or Gamification. He works for Serious Games Solutions, a division of Promotion Software GmbH, the software developer behind the "Emergency Lernspiel" (Learning game). He developed games or gamified apps for mobile platforms, browser and HoloLens. His focus is on game design and project management.

Korbinian Moeller received his PhD in Psychology from University of Tübingen, Germany, in 2010. Currently he is Professor of Mathematical Cognition at Loughborough University, United Kingdom. His research interests include neurocognitive foundations of mathematical cognition and developmental psychology in the context of educational games and learning analytics.

Manuel Ninaus received his PhD in Psychology from University of Graz, Austria, in 2015. Currently he is PostDoc at the University of Innsbruck, Austria. He is elected board member of the Serious Games Society. His research interests include neuroscience and educational psychology in the context of educational games and learning analytics.

Christian Scharinger Christian Scharinger received the PhD degree in cognitive science from the University of Tuebingen in 2015. He currently works at the Multimodal Interaction Lab of the Knowledge Media Research Center Tuebingen. He has a profound expertise in (neuro-) physiological measures like eye-tracking and EEG. In his research he tries to combine basic and applied research areas. His research interests comprise working memory, executive functions, learning, hypertext reading, web searching, and multimedia, with a strong focus on physiological measures of cognitive load.

Natalia Sevcenko Natalia Sevcenko is currently pursuing her doctorate in psychology at the LEAD Graduate College and Research Network at the University of Tübingen. She received her bachelor's and master's degrees in psychology in 2015 and 2017 respectively at the Eberhard-Karl-University of Tübingen. Her research interests lie in the field of human-machine interaction, especially in the area of behavior and sensor-based measurement of cognitive states of operators and their relation to learning outcomes and personal predisposition.

Franz Wortha Franz Wortha is currently pursuing his Ph.D. in Psycholgoy at the LEAD Graduate School and Research Network at the University of Tübingen. He received his Bachelor's degree in Industrial Engineering from the University of Applied Sciences in Dresden in 2013 and his Master's degree in Psychology from Technische Universität Dresden in 2016. His research interests lie in self-regulated learning with a focus on metacognitive and emotional processes and their relation to learning outcomes and personal predisposition.

Enkelejda Kasneci is a Professor of Computer Science at the University of Tübingen, Germany, where she leads the Human-Computer Interaction Lab. As a BOSCH scholar, she received her M.Sc. degree in Computer Science from the University of Stuttgart in 2007. In 2013, she received her PhD in Computer Science from the University of Tübingen. For her PhD research, she was awarded the research prize of the Federation Südwestmetall in 2014. From 2013 to 2015, she was a postdoctoral researcher and a Margarete-von-Wrangell Fellow at the University of Tübingen. Her research evolves around the application of machine learning for intelligent and perceptual human-computer interaction. She serves as academic editor for PlosOne and as a TPC member and reviewer for several major conferences and journals.