Predicting Cognitive Load in an Emergency Simulation Based on Behavioral and Physiological Measures

Tobias Appel
tobias.appel@uni-tuebingen.de
LEAD Graduate School and Research Network
Tübingen, Germany

Natalia Sevcenko
Leibniz-Institut für Wissensmedien
Tübingen, Germany

Franz Wortha
LEAD Graduate School and Research Network
Tübingen, Germany

Katerina Tsarava
Leibniz-Institut für Wissensmedien
Tübingen, Germany

Korbinian Moeller
Leibniz-Institut für Wissensmedien
Tübingen, Germany

Manuel Ninaus
Leibniz-Institut für Wissensmedien
Tübingen, Germany

Enkelejda Kasneci
University of Tübingen
Tübingen, Germany

Peter Gerjets
Leibniz-Institut für Wissensmedien
Tübingen, Germany

ABSTRACT

The reliable estimation of cognitive load is an integral step towards real-time adaptivity of learning or gaming environments. We introduce a novel and robust machine learning method for cognitive load assessment based on behavioral and physiological measures in a combined within- and cross-participant approach. 47 participants completed different scenarios of a commercially available emergency personnel simulation game realizing several levels of difficulty based on cognitive load. Using interaction metrics, pupil dilation, eye-fixation behavior, and heart rate data, we trained individual, participant-specific forests of extremely randomized trees differentiating between low and high cognitive load. We achieved an average classification accuracy of 72%. We then apply these participant-specific classifiers in a novel way, using similarity between participants, normalization, and relative importance of individual features to successfully achieve the same level of classification accuracy in cross-participant classification. These results indicate that a combination of behavioral and physiological indicators allows for reliable prediction of cognitive load in an emergency simulation game, opening up new avenues for adaptivity and interaction.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; • **Computing methodologies** → *Cross-validation.*

KEYWORDS

Cognitive Load, Eye Tracking, Heart Rate, Multimodal, Classification

ACM Reference Format:
Tobias Appel, Natalia Sevcenko, Franz Wortha, Katerina Tsarava, Korbinian Moeller, Manuel Ninaus, Enkelejda Kasneci, and Peter Gerjets. 2019. Predicting Cognitive Load in an Emergency Simulation Based on Behavioral and Physiological Measures. In *2019 International Conference on Multimodal Interaction (ICMI '19), October 14–18, 2019, Suzhou, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3340555.3353735

ICMI '19, October 14–18, 2019, Suzhou, China

ACM ISBN 978-1-4503-6860-5/19/10…$15.00
https://doi.org/10.1145/3340555.3353735

1 INTRODUCTION

Real-time user modeling in general and cognitive modeling in particular are essential for successfully implementing adaptive interfaces and environments. One important aspect of cognitive modeling is its consideration of cognitive load, which refers to the degree to which cognitive resources such as working memory are recruited while performing a task [6, 10]. Often it is beneficial to adapt a system in a way that the cognitive load experienced by a user does not exceed a critical level. An e-learning environment, for example, should

neither provide trivial learning materials, nor overstrain the user with materials and tasks they cannot cope with [30]. This also holds true for most computer environments involving different levels of difficulty, such as games, training simulations, or tutoring systems. An example of successful implementation of this kind of adaptivity is demonstrated by Yuksel and colleagues [35]. They used EEG data for adapting difficulty during piano lessons, which they observed to increase participants' learning gains. Wilson and Russel used physiological measures such as EEG, respiration, and heart rate, but also eye-fixation behavior to realize adaptivity [34] in an aviation simulation. The provided real-time adaptive feedback enhanced participants' performance.

However, assessing cognitive load can not only be used for adaptation. In many situations, its assessment can be valuable on its own, in particular in the absence of other outcome measures. In this case, reliable indicators of cognitive load may allow for evaluating which interaction modality is easiest to use, which type of interface causes the desired degree of cognitive load, or how difficult a certain activity is [3].

Yet many traditional ways of assessing cognitive load are impractical for many situations. Using questionnaires like the NASA task-load index (TLX) [15] make the user aware of the assessment, interrupting and impairing task performance and reducing immersion. As a retrospective measure, this is not an issue, but for real-time measuring and adaptation it is not the right tool.

Other methods make use of physiological changes caused by cognitive load to derive respective indicators. Direct impact of cognitive load was suggested to be measurable considering brain activity (e.g., by means of EEG) and participant-specific predictions were found to be accurate [22, 24]. Methods for measuring brain activity such as EEG have the drawback of usually being intrusive, uncomfortable to use, and requiring a high expertise to be set up. While there are mobile versions of such devices, mitigating some of the drawbacks, the intrusive nature making users aware of being monitored still remains. There are, however, physiological parameters which can be measured indirectly or with little effort. Eye-fixation behavior is a prime example of such measures because it is easy to measure, less invasive for the user, and allows for using changes in pupil diameter as an indicator of cognitive load [1, 20]. Moreover, changes in heart rate are also a common effect of variations of cognitive load that can be measured with little effort and intrusion. Combining these two streams of data increases predictive power [31].

In addition to estimating cognitive load through measure of pupil dilation, eye tracking has the benefit of allowing for evaluating participants' gaze behavior. This allows insights into cognitive processes and offers behavioral data as another means for assessing cognitive load. Gaze information can be used in conjunction with interaction log-files to better grasp how participants react under low or high cognitive load.

The advantage of jointly considering several (physiological) data streams for the assessment of cognitive load was previously demonstrated in several studies [13, 16, 17, 28]. Hussain et al. used a combination of heart rate, skin conductance, respiration data, and eye-tracking [17]. In addition, they synchronized these physiological measures with facial features and behavioral data in pursuit of classifying cognitive load under the effect of affective interference. Haapalainen and colleagues further added a skin temperature sensor and an EEG-headset to this set of sensors and achieved a mean classification rate of 81.1% for the distinction between low and high cognitive load for 6 cognitive tasks [13].

A frequent problem in practice arising from the estimation of cognitive load being participant-specific results in a lack of generalizability [23]. As a consequence, a cognitive model for each participant needs to be trained in order for adaptive systems to work. This would necessitate a lengthy calibration period involving examples of low and high cognitive load for each participant. This often seems impractical and may deter users from actually using the system in the first place.

In this article, we combine behavioral and physiological measures in a novel multimodal approach for classifying cognitive load in an emergency simulation game. We specifically aim to increase cross-participant generalizability up to the point where the accuracy of cross-participant classification is the same as within-participant classification. In particular, we want to develop a method for classifying cognitive load that satisfies the criteria of i) allowing for a robust estimation with ii) high classification accuracy and iii) generalizability across participants for potential iv) real-time application.

2 SETUP AND STIMULI

Task

Participants had to perform in different scenarios taken from an adapted version of the real-time simulation game Emergency [12]. The three scenarios were adapted specifically for the purpose of this study, Figure 1 providing an example. The goal of the game is to coordinate emergency forces. Participants take control of paramedics, ambulances, and firefighters to save people from emergency situations and fight fires in the scenario.

Participants first completed a tutorial introducing all game mechanics with instructions and in absence of time pressure, followed by three different scenarios: a car crash, burning buildings, and a train crash. Each scenario was presented in three versions, "easy", "medium", and "hard". The scenarios were always presented in the same order, starting with the easiest version of the car crash scenario going through the hardest version of the train crash scenario. The simulation's

difficulty was raised by increasing the number of tasks a participant had to complete as well as the number of units available, while maintaining the same time constraint. This necessitates planning more steps ahead, while also requiring more micro-managing and better prioritizing. As planning gets more sophisticated and time pressure increases, cognitive load should increase as a consequence. Because we lack means of measuring cognitive load directly, we instead aimed to classify task difficulty as a proxy for cognitive load. All instructions were presented in German.

Scenario 0: Tutorial. The tutorial provided instructions on how to give orders to emergency forces and which tasks needed to be carried out for successfully completing the game. It also introduced the different units participants needed to coordinate and their purpose in the simulation. There was no time limit.

Scenario 1: Car Crash. The first scenario featured a car crash at a crossroads. Some accident victims were trapped within their cars and the participant needed to send firefighters to free them. At the same time, other victims needed treatment by a paramedic before being transported to the hospital. There was a 5 minute time limit for this scenario.

Scenario 2: Burning Buildings. In this scenario, several buildings were on fire. This poses a more dynamic threat as the fire can spread to neighbouring buildings and paramedics cannot operate in close proximity to fire. Several victims needed to be saved from burning buildings with ladders, while others were in need of medical aid. To extinguish the flames, fire trucks and firefighters could be used, which differ in their effectiveness. The dynamics of the fire made this scenario inherently more difficult than the first one. The time limit was 7 minutes and 30 seconds.

Scenario 3: Train Crash. In the final scenario, participants faced a derailed train that had crashed into a building. As a consequence of the train having hit a building, the building had caught on fire. This also threatened the surrounding buildings. Furthermore, train wagons were deformed, making it necessary for them to be cut open to save trapped victims. Adding to these tasks, there were victims in need of medical aid. By combining all challenges participants faced before, the final scenario further increased the difficulty compared to the previous ones. This especially emphasizes the need for prioritizing actions because firefighters could either cut victims free or extinguish fires. Taking into consideration the complexity and volume of this scenario, the time limit was set to 10 minutes.

Apparatus

The experiment took place in individual sessions in a laboratory setting under constant light conditions. The emergency simulation and the eye tracker were installed on a notebook with a 16" screen driven at 1920 x 1080 resolution. The position of each participant was individually determined based on the calibration of the eye tracker. No chin rest was used. For the recording of eye-fixation behavior, we used a RED250 eye tracker from SensoMotoric Instruments (SMI) in combination with the SMI Experiment Center 3.7.60 software. The eye tracker was calibrated using SMI's integrated 9-point calibration procedure.

For the measurement of heart rate, a custom-made Bitalino wearable was used in combination with OpenSignals Revolution software. To increase the signal-to-noise ratio, we attached three pre-gelled electrodes to participants' chests, which were cable-connected with the Bitalino unit. For technical reasons, an additional laptop was necessary to acquire heart-rate data. The laptop was connected to the measuring computer via Bluetooth. The whole setup is depicted in Figure 2.

Participants

We gathered data from 47 right-handed participants (*mean age* = 24.6, *SD* = 6.3, 33 females). None of them reported psychological, neurological, or cardiovascular diseases. All participants provided written informed consent and were able to speak German on native speaker level. Participants were informed about the study aims and received monetary compensation after the experiment. We had to exclude 7 participant due to problems with recordings of eye-tracking, physiological or meta-data. Another 2 were excluded because they reported that they did not take the experiment seriously. Finally, 2 participants did not provide enough usable data for all scenarios, rendering their data partly unusable. The data of all remaining 36 participants was considered in the analyses. Participants with periods of low tracking ratios – time spans with invalid data caused by the pupil not being detected reliably – were deliberately included in our dataset. Including them renders the dataset more realistic and consequently the results more accurately reflect a real-world scenario.

3 FEATURES

We looked at features which have been used successfully in the past to detect cognitive load. We grouped the features into different categories of i) pupil dilation, ii) fixations, iii) saccades, iv) heart rate, and v) in-game activity. In this section, we describe them in more detail.

Pupil

Many studies have successfully used pupil diameter as an indicator for cognitive load [7, 19, 25]. Increasing cognitive load was observed to decrease parasympathetic activity in the peripheral nervous system, leading to an increase in pupil

Figure 1: Example of how the task looked like.

Figure 2: Setup that was used during data collection.

diameter [20]. This effect was found for different tasks, including short-term memory, language processing, reasoning, perception, as well as sustained and selective attention [1]. The effect was also consistently observed within a task, between tasks, and between individuals [18]. In a preprocessing step, we first removed all data points where pupil diameter was 0 or negative. Such artifacts typically occur in case of invalid data points. We also removed data points up to 100 ms directly before and after a blink. During these periods, the pupil is partly occluded by the eyelid or eyelashes and cannot be detected reliably. Finally, we linearly interpolated small gaps of up to 50 ms to increase the amount of usable data. Because the analyzed time periods are very short and thus susceptible to noise, we used the median of pupil diameter instead of the more commonly used mean. We expect this to result in more robust features and more reliable predictions. Additionally, we used the pupil diameter maximum to also consider peaks in pupil diameter. Both parameters are expected to increase with increasing task difficulty.

Fixations

Fixations describe a stable gaze on the same location usually lasting between 200 ms and 350 ms [27]. The number of fixations per second depends on many factors. Higher cognitive load was found to lead to fewer but longer fixations [8], whereas time pressure tended to decrease fixation duration while increasing the number of fixations per second [33]. We used the number of fixations per second as a feature. The difficulty levels of our scenarios were mainly driven by the number of emergency personnel to coordinate, the number of sub-tasks to perform and increases in time pressure, leading to the expectation of an increase in the number of fixations per second.

Saccades

Rapid eye movements that usually occur between fixations are called saccades. How cognitive load and task difficulty influence saccade characteristics strongly depends on the task at hand. There is evidence pointing towards an increase of average saccade amplitude in search tasks compared to free viewing [32]. With increasing task difficulty, the amount of visual exploration should decrease and participants' gaze

behavior should be dominated by specific goal-directed saccades. Hence, we expect to find higher saccade frequency during high-difficulty tasks. Because participants have to navigate the interface efficiently to perform well during the simulation, we expect a higher saccade amplitude during phases of high task difficulty.

Another form of saccades are so-called microsaccades. Microsaccades are small involuntary eye movements that can occur during a fixation. Studies have tied them to cognitive load in different situations. Non-visual tasks appear to reduce the number of microsaccades [9, 21, 29], while visually more demanding tasks appear to increase the frequency of microsaccades [2]. We used the method suggested by Krejtz and colleagues [21] to detect microsaccades, but focused on microsaccade frequency rather than amplitude or velocity. We expected to see an increase of miscrosaccade frequency with task difficulty.

Heart Rate

Just like pupil diameter increases as a consequence of reduced parasympathetic activity, heart rate was observed to increase as well. Various studies investigated the relationship between cognitive load and heart rate [4] and substantiated this association.

We used the raw ECG signal recorded by the Bitalino unit and processed it with the python package biosppy [5] which uses an approach by Hamilton [14] for QRS detection and provides us with a timestamp for every R-peak. A QRS complex is the main spike in an ECG signal, marking a heart beat, and an R-peak is the highest point of this complex. Using the exact R-peak points, we then calculated the number of heart beats per second, which we used as a feature for assessing task difficulty.

In-game Activity

Not only physiological measures can be used to assess task difficulty, but also the way participants interact with the simulation. Periods of high task difficulty should be accompanied by higher in-game activity. Successfully completing different sub-tasks – such as putting out fires or saving victims under time pressure – requires coordinating emergency forces while keeping the overall setting in mind. [1]

We used actions per second as a measure for in-game activity. Actions comprise opening the menu of an emergency respondent, selecting a specific command, and selecting a target for that command. For instance, commanding a paramedic to care for a certain victim results in 3 actions.

[1]Using the number of clicks may seem to render the classification problem trivial, but even without the in-game activity as a feature, classification accuracy is still around 70% and as our results show, it is not the dominant feature.

4 METHOD

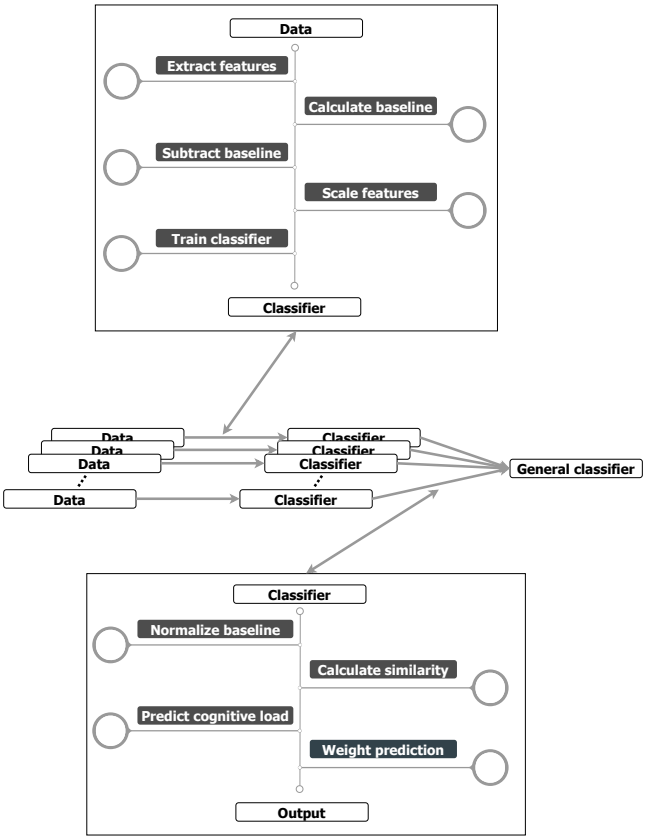

Figure 3: Schematic overview of our method showing how we train individual classifiers and combine them to a general one.

Our goal was to develop a robust classifier that works reliably as well as independently of individual participants. The main rationale underlying our approach was to train classifiers based on participants from a database and apply these classifiers to novel participants.

Participant-specific Classifiers

In the within-participant part of our approach, we aimed to train a classifier for each participant and scenario, classifying whether a participant currently plays the easy or hard version of the simulation. To train such classifiers, we needed samples from low- and high-difficulty periods to learn from. This prerequisite was achieved by drawing 60 random samples of 4 s intervals, half of which are obtained from the easy versions of each scenario and the other half from the hard version. This process was repeated for each participant and scenario. The 4 s interval enabled us to sample without overlap and still include participants that had completed the scenario well before the time limit was reached. Furthermore,

this interval length enabled us to reject samples with more than 30 % missing data, while still including all participants. This prevented any bias that may have been introduced by systematically excluding well-performing participants.

For each sample obtained, we calculated the features mentioned in Section 3. We used the tutorial as a baseline, meaning that we extracted the same features for the whole tutorial and used them for standardization, resulting in features that were a percentage change of the baseline. This did not require any further processing because all the used features were standardized per second. Choosing the tutorial as a baseline made it unnecessary to employ a dedicated calibration or baseline period, which increases real-world applicability of our approach.

After extracting the features, we z-standardized all features for each scenario. This resulted in all features having a mean of 0 and a standard deviation of 1, ensuring that all features had the same normalized scale and that any machine learning algorithm did not weight features according to their scaling. Finally, we trained a forest of 100 extremely randomized trees (Extra-Trees) [11] with a single participant's data to obtain a classifier specific to that participant. This resulted in one classifier per participant and scenario being able to distinguish a sample as either originating from a period of high or low task difficulty. Instead of just classifying into "low" or "high", we used class probabilities, which are scores between 0 and 1. All outputs above 0.5 were considered to reflect high task difficulty and all below 0.5 were considered to indicate low task difficulty. This provided the advantage of incorporating confidence into the prediction, improving generalizability in the following step. Our method was implemented in Python using the scikit-learn toolbox [26].

General Classifiers

Based on the classifiers for each participant, the next step was to generalize across participants. A naive way of approaching this issue would be to either train one classifier for all participants or to blindly apply the trained classifiers to other participants. It does, however, make little sense to apply participant A's classifier to participant B if their physiological or behavioral features are very different. Therefore, we weighted the prediction of individual classifiers according to how similar they were to the participant we wanted to apply them to.

First, we standardized the baselines for participant, guaranteeing that certain baseline features do not receive a higher weight when calculating the Euclidean distance between two baselines. Features on a larger scale tend to dominate the distance, because they result in larger numbers (e.g., there are a lot more fixations per second than in-game actions).

Let x be a novel participant whose cognitive load we want to classify, $sample_x$ a sample of x characterized with a set of features, and Y the set of participants we trained on. Every $y \in Y$ has a classifier c_y that predicts a value between 0 and 1 for $sample_x$. We combine these predictions according to the following equations:

$$sim(x, y) = \frac{1}{\sum weights_{c_y} |baseline_x - baseline_y|}$$

$$pred(sample_x) = \frac{\sum_{y \in Y} sim(x, y) pred_{c_y}(sample_x)}{\sum_{y \in Y} sim(x, y)}$$

$sim(x, y)$ refers to the baseline similarity between participants x and y, $weights_{c_y}$ to the normalized feature weights of classifier c_y, and $pred_y$ to the prediction of classifier c_y. This means we let each classifier c_y predict the cognitive load and weight these predictions according to the similarity between participants x and y. Additionally, we factor the feature weights of classifier c_y into the similarity, giving a higher weight to more important features. Dividing by the sum of all similarities normalizes these similarities and ensures that the prediction's final result is within the interval of $[0, 1]$.

5 RESULTS

In order to provide a frame of reference for the accuracy of our method, we first present descriptive statistics for the 3 scenarios in Table 1. This illustrates how difficult the individual scenarios were. The higher the increase in difficulty, the larger differences between the easy and the hard version should be and consequently the performance of our classification should be better. We focus on the distinction between easy and hard task difficulty, but our method can also be applied to the classification problems "easy vs. medium", "medium vs. hard" and for the multi-class problem "easy vs. medium vs. hard". Evaluating these problems and their results, however, is beyond the scope of this article.

Table 1: Descriptive statistics about the difficulty of the scenarios.

scenario	finished on time	average time of completion, if finished
1 easy	83.33 %	4:12
1 hard	83.33 %	4:16
2 easy	97.22 %	3:58
2 hard	36.11 %	6:31
3 easy	97.22 %	7:01
3 hard	33.33 %	8:57

Participant-specific Results

To evaluate how well our cross-participant approach works, we first ran participant-specific classifiers. This was performed using 10-fold cross-validation, to ensure that we do not artificially increase classification accuracy by overfitting. Table 2 shows the average accuracy for each scenario.

Table 2: Mean accuracy for within-participant classification of cognitive load.

scenario	accuracy
1	79.03%
2	70.14%
3	72.13%

Feature Weights

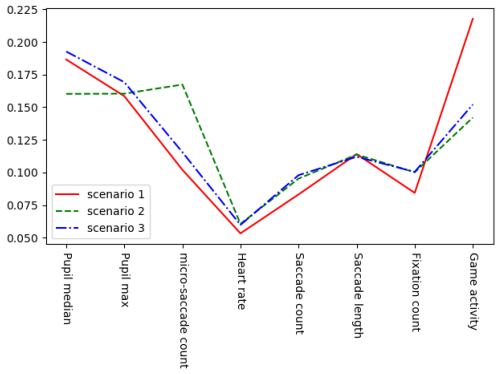

Figure 4: Average feature importance for each of the 3 scenarios.

To better understand our results, we also present average feature weights in Figure 4. Interestingly, pupil features have the highest weight, making them the most important features. We expected a large contribution from the pupil data and this result is in line with the literature. Another important feature is in-game activity. In particular, it is very important in the first scenario, because participants are not yet familiar with the controls, but by the time they have to perform the hard version they already acquired more experience, increasing the difference in in-game activity. Learning and experience reduce this effect in the subsequent scenarios. Microsaccades are the next most important category. Several publications indicated that cognitive load has an effect on microsaccade frequency [21], but usually with higher relative influence. Our recording frequency of 250 Hz may be too low to detect all microsaccades and most studies recorded microsaccades during very long fixations of several seconds, whereas our task resulted in rather short fixations.

As expected, saccade and fixation characteristics are predictive for task difficulty, but do not carry as much weight as other features. Finally, heart rate does not seem to matter very much for our algorithm. Nevertheless, looking at mean values per difficulty level, we still consistently find heart rate to increase with difficulty. A possible reason might be the short intervals analyzed. 4 s may be a very short time frame to evaluate heart rate meaningfully.

Cross-participant Results

After we trained our participant-specific classifiers, we applied them for cross-participant classification according to the schema described in Section 4. We performed leave-one-out cross-validation on the participant level, meaning that we withheld the participant to whom we applied our cross-participant algorithm. To also test how well our approach performs across different situations, we used classifiers which had been trained either on the same scenario or on one of the two others. The average results for classification accuracy are shown in Table 3.

Table 3: Mean accuracy for cross-participant classification of cognitive load when applying classifiers trained on different scenarios.

Classifier from \ applied to	1	2	3
1	80.56%	67.92%	70.10%
2	73.70%	70.37%	67.04%
3	76.13%	67.73%	69.81%

It is noticeable that our approach does not lose accuracy performing cross-subject classification compared to within-subject classification. Only in the case of scenario 3, accuracy drops slightly. This is likely to be the case, because towards the end of our experiment subjects adapt strategies that differ between them. Overall our approach generalizes well on participant level, which can be attributed to the weighting and various forms of standardization. We performed the same classification with only the 10 most similar participants instead of all participants and still got the same accuracy. This may be used to save runtime in case little processing power is available or when the number of participants gets very large.

Also note that the accuracy barely drops when we use data from another scenario, so our method does not only generalize well across participants but also across different scenarios. This is signified by the columns of Table 3. The diagonal shows the the results obtained by training and evaluating on the same scenario setting the bar for other classifiers, whereas the rest of the column depicts results from sub-optimal classifiers trained on other scenarios. As the

scenarios differ in regards to how well easy and hard version can be distinguished, meaningful row-wise comparisons can not be made.

With regard to runtime, calculation of features took 0.699 ms on average and weighted classification took 6.723 ms, thus resulting in a total runtime of 7.422 ms. The reported performance was measured on a laptop with an Intel(R) Core(TM) i7-7700HQ with 16 GB RAM running a non-optimized version of our algorithm that does not make use of parallel processing. Limiting the number of participants we consider for classification to the n most similar ones would further speed up our algorithm while still maintaining a high accuracy.

6 DISCUSSION

Our goal was a i) robust estimator offering ii) high accuracy, iii) generalizing well across participants, and is also iv) real-time capable. In conclusion, our approach satisfies all these criteria.

Firstly, we showed the robustness of our approach. We operate with a suboptimal baseline derived from the tutorial and did not exclude participants with low tracking rates. Furthermore, we use only one baseline from the very beginning of the experiment. As a consequence, fatigue influences pupil data over the course of the experiment, decreasing its diameter and thereby counteracting some effects of changes in cognitive load. Furthermore, luminance changes caused by the dynamic nature of the simulation also add noise that our methods shows resilience towards.

Secondly, a mean classification accuracy of 72 % appears sufficient for most real-world applications. Actual classification accuracy may most likely be higher because we use task difficulties for classification and not actual cognitive load. During a heterogeneous task, cognitive load is usually not at a constant level and can be low within a difficult task high within an easy task. Moreover, for participants that finished on time, we noticed a drop in predicted task difficulty towards the end of the task indicating they may not be challenged anymore. Additionally, most participants started any version of a scenario with low predicted task difficulty, likely because they were adjusting to the task. Both of these circumstances reduce nominal accuracy of our approach even though predictions may be accurate.

Thirdly, our results show that our method is able to generalize across participants. There is no drop in accuracy when we apply weighted predictions to withheld participants, indicating that we can expect the same level of accuracy when we apply our method to new participants. This even holds true when we restrict training the algorithm to the 10 most similar participants.

This is relevant for the last criterion of being executed in real-time. When processing power is limited, the number of participants whose classifiers are applied can be restricted, reducing runtime to a fraction of its original time without loss of classification accuracy. This will allow most devices to run the algorithm at an acceptable frequency.

Finally, our method maintaining its high classification accuracy even when applied across different scenarios indicates an additional degree of robustness.

Apart from the ability to generalize there are other benefits to the presented method. For instance, converting the continuous output to a binary classification is not mandatory, as the output can be used directly. The higher the score, the more likely is cognitive load to also be high. In the case of an adaptive environment, task difficulty could be adjusted in case it is found, for example, to be below 0.3 or above 0.7.

There are, however, limitations to our approach. We standardized features of each participant to a mean of 0 and a standard deviation of 1, which may cause problems when done in real time. If we assume the same order of tasks as in this work, participants start with an easy version of the first scenario. This means we only have data from periods of low cognitive load during this task and scaling does not work as expected. This may partly be mitigated by applying the scaling function from participants from our database, meaning we subtract the mean of a recorded participant and divide by their standard deviation. This will still cause minor loss in accuracy, but partly solves the problem. As soon as we have data from periods of low and high cognitive load, this is not an issue anymore.

While our approach can be used for many different tasks, the trained estimators can not. The fixation and saccade characteristics are specific to our simulation, so any classifier we trained has limited use for other tasks. If we ignore fixations and saccades to focus on the less task-dependent features like pupil diameter, microsaccades and heart rate, the range of application widens, but the accuracy for a specific task is reduced.

As a future perspective, we are planning a follow-up study using the method presented in this article to create an adaptive version of the emergency simulation. The 36 participant-specific classifiers we trained should suffice as a database to evaluate cognitive load of new participants in real time and we are confident that we can implement a system allowing for real-time adaptation of cognitive load caused by the simulation.

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 was a doctoral student of the LEAD Graduate School and Research Network.

REFERENCES

[1] Jackson Beatty. 1982. Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin* 91, 2 (1982), 276.

[2] Simone Benedetto, Marco Pedrotti, and Bruce Bridgeman. 2011. Microsaccades and exploratory saccades in a naturalistic environment. *Journal of Eye Movement Research* 4, 2 (2011), 1–10.

[3] Maneesh Bilalpur, Mohan Kankanhalli, Stefan Winkler, and Ramanathan Subramanian. 2018. Eeg-based evaluation of cognitive workload induced by acoustic parameters for data sonification. In *Proceedings of the 2018 on International Conference on Multimodal Interaction*. ACM, 315–323.

[4] Yati N Boutcher and Stephen H Boutcher. 2006. Cardiovascular response to Stroop: effect of verbal response and task difficulty. *Biological Psychology* 73, 3 (2006), 235–241.

[5] Carlos Carreiras. 2015–. BioSPPy: Biosignal Processing in Python. https://github.com/PIA-Group/BioSPPy

[6] Fang Chen, Jianlong Zhou, Yang Wang, Kun Yu, Syed Z Arshad, Ahmad Khawaji, and Dan Conway. 2016. *Robust multimodal cognitive load measurement*. Springer.

[7] Siyuan Chen and Julien Epps. 2013. Automatic classification of eye activity for cognitive load measurement with emotion interference. *Computer Methods and Programs in Biomedicine* 110, 2 (2013), 111–124.

[8] Michel De Rivecourt, Marianne Kuperus, Wendy J Post, and Lambertus JM Mulder. 2008. Cardiovascular and eye activity measures as indices for momentary changes in mental effort during simulated flight. *Ergonomics* 51, 9 (2008), 1295–1319.

[9] Xin Gao, Hongmei Yan, and Hong-jin Sun. 2015. Modulation of microsaccade rate by task difficulty revealed through between-and within-trial comparisons. *Journal of Vision* 15, 3 (2015), 3–3.

[10] Peter Gerjets, Carina Walter, Wolfgang Rosenstiel, Martin Bogdan, and Thorsten O. Zander. 2014. 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* 8 (2014), 385. https://doi.org/10.3389/fnins.2014.00385

[11] Pierre Geurts, Damien Ernst, and Louis Wehenkel. 2006. Extremely randomized trees. *Machine Learning* 63, 1 (2006), 3–42.

[12] Promotion Software GmbH. 1999. World of Emergency. https://www.world-of-emergency.com/?lang=en

[13] Eija Haapalainen, SeungJun Kim, Jodi F Forlizzi, and Anind K Dey. 2010. Psycho-physiological measures for assessing cognitive load. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*. ACM, 301–310.

[14] Patrick S Hamilton. 2002. Open source ECG analysis software documentation. *Computers in Cardiology* 2002 (2002), 101–104.

[15] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in Psychology*. Vol. 52. Elsevier, 139–183.

[16] SM Hadi Hosseini, Jennifer L Bruno, Joseph M Baker, Andrew Gundran, Lene K Harbott, J Christian Gerdes, and Allan L Reiss. 2017. Neural, physiological, and behavioral correlates of visuomotor cognitive load. *Scientific Reports* 7, 1 (2017), 8866.

[17] M Sazzad Hussain, Rafael A Calvo, and Fang Chen. 2013. Automatic cognitive load detection from face, physiology, task performance and fusion during affective interference. *Interacting with Computers* 26, 3 (2013), 256–268.

[18] Daniel Kahneman. 1973. *Attention and effort*. Vol. 1063. Citeseer.

[19] Jeffrey M Klingner. 2010. *Measuring cognitive load during visual tasks by combining pupillometry and eye tracking*. Ph.D. Dissertation. Stanford University Stanford, CA.

[20] Arthur F Kramer. 1991. Physiological metrics of mental workload: A review of recent progress. *Multiple-task Performance* (1991), 279–328.

[21] Krzysztof Krejtz, Andrew T Duchowski, Anna Niedzielska, Cezary Biele, and Izabela Krejtz. 2018. Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze. *PLOS ONE* 13, 9 (2018), e0203629.

[22] Shiba Kuanar, Vassilis Athitsos, Nityananda Pradhan, Arabinda Mishra, and Kamisetty R Rao. 2018. Cognitive Analysis of Working Memory Load from EEG, by a Deep Recurrent Neural Network. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2576–2580.

[23] Jesus L Lobo, Javier D Ser, Flavia De Simone, Roberta Presta, Simona Collina, and Zdenek Moravek. 2016. Cognitive workload classification using eye-tracking and EEG data. In *Proceedings of the International Conference on Human-Computer Interaction in Aerospace*. ACM, 16.

[24] Caitlin Mills, Igor Fridman, Walid Soussou, Disha Waghray, Andrew M Olney, and Sidney K D'Mello. 2017. Put your thinking cap on: detecting cognitive load using EEG during learning. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*. ACM, 80–89.

[25] Oskar Palinko, Andrew L Kun, Alexander Shyrokov, and Peter Heeman. 2010. Estimating cognitive load using remote eye tracking in a driving simulator. In *Proceedings of the 2010 Symposium on Eye-tracking Research & Applications*. ACM, 141–144.

[26] 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. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[27] Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin* 124, 3 (1998), 372.

[28] Christian Scharinger, Yvonne Kammerer, and Peter Gerjets. 2015. Pupil Dilation and EEG Alpha Frequency Band Power Reveal Load on Executive Functions for Link-Selection Processes during Text Reading. *PLOS ONE* 10, 6 (06 2015), 1–24. https://doi.org/10.1371/journal.pone.0130608

[29] Eva Siegenthaler, Francisco M Costela, Michael B McCamy, Leandro L Di Stasi, Jorge Otero-Millan, Andreas Sonderegger, Rudolf Groner, Stephen Macknik, and Susana Martinez-Conde. 2014. Task difficulty in mental arithmetic affects microsaccadic rates and magnitudes. *European Journal of Neuroscience* 39, 2 (2014), 287–294.

[30] Martin Spüler, Carina Walter, Wolfgang Rosenstiel, Peter Gerjets, Korbinian Moeller, and Elise Klein. 2016. EEG-based prediction of cognitive workload induced by arithmetic: a step towards online adaptation in numerical learning. *ZDM* 48, 3 (01 Jun 2016), 267–278. https://doi.org/10.1007/s11858-015-0754-8

[31] Zach Chuanzhong Tan, Bryan Reimer, Bruce Mehler, and Joseph F Coughlin. 2011. Detection of elevated states of cognitive demand in drivers in a naturalistic driving environment. In *Proceedings of the 2nd Annual International Conference on Advanced Topics in Artificial Intelligence (ATAI 2011)*. 24–25.

[32] Benjamin W Tatler, Roland J Baddeley, and Benjamin T Vincent. 2006. The long and the short of it: Spatial statistics at fixation vary with saccade amplitude and task. *Vision Research* 46, 12 (2006), 1857–1862.

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

[34] Glenn F Wilson and Christopher A Russell. 2007. Performance enhancement in an uninhabited air vehicle task using psychophysiologically determined adaptive aiding. *Human Factors* 49, 6 (2007), 1005–1018.

[35] Beste F Yuksel, Kurt B Oleson, Lane Harrison, Evan M Peck, Daniel Afergan, Remco Chang, and Robert JK Jacob. 2016. 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, 5372–5384.