Student Conference Lübeck

Proceedings Article

Examining Software Developers' Cognitive Load During Daily Activities with Wearables

Annemarie Uhlig [a,*] · Charlotte Brandebusemeyer [b] · Fabian Stolp [b] · Hawzhin Hozhabr Pour [c] · Bert Arnrich [b]

[a] Student of Medical Informatics, Universität zu Lübeck, Lübeck, Germany
[b] Hasso Plattner Institute, University of Potsdam, Potsdam, Germany
[c] Institute of Medical Informatics, Universität zu Lübeck, Lübeck, Germany
[*] Corresponding author, email: annemarie.uhlig@student.uni-luebeck.de; hawzhin.hozhabrpour@uni-luebeck.de

Received 04 February 2025; Accepted 24 June 2025; Published online 15 July 2025

Abstract

In the dynamic field of software development, deadline-driven projects and the constant demand for innovation often result in an underestimated burden: cognitive load. Traditional methods, e.g., questionnaires, capture the subjectively perceived cognitive load in a time- and situation-dependent manner. However, wearable sensors could provide an objective and continuous evaluation of cognitive load in everyday work situations. Our exploratory data analysis employs a holistic approach, using recorded multi-modal physiological signals during software development-typical tasks from four participants who also filled in a NASA TLX questionnaire. Following data cleaning and a comprehensive analysis, a relation between perceived cognitive load and psychophysiological data was identified using metrics such as event-related synchronization.

I. Introduction

Managing mental resources and the associated cognitive load is crucial for productivity and solving tasks in today's software development. Cognitive load theory defines cognitive load as the amount of information the working memory must process at any given time [1]. High cognitive load negatively impacts an individual's productivity and learning process [1]. Cognitive load increases, e.g., with daily, mentally demanding activities [2].

Software developers handle various tasks, including writing code, debugging, code documentation, and writing emails. Often, cognitive load is measured using questionnaires. However, questionnaires can interrupt work routines, and answers are time- and situation-dependent and subjective. Consequently, some studies assessed the cognitive load by measuring physiological data and linking them to cognitive load [2, 3]. Using sensors could potentially offer an objective method to assess processing load and estimate cognitive load as a continuous measurement in real time. Wearables equipped with sensors that measure physiological signals, such as electrodermal activity (EDA), electroencephalography (EEG), and pupil dilation, can be used to derive cognitive load.

Electrodermal Activity (EDA): EDA measures the influence of the central and sympathetic nervous system on sweat glands through changes in skin resistance or electrical potential [4]. It is an index for directly investigating stress-related effects on bodily functions and can describe cognitive load levels [2]. The signal consists of a tonic and a phasic component [2]. The tonic component refers to changes in the skin conductance level (SCL), which gradually changes over time [2]. The phasic component describes responses in the phasic skin conduction (SCR) [5]. It is a rapidly changing signal in response to a stimulus [5]. A high cognitive load induces a physiological reaction from the sympathetic nervous system in terms of increased skin conductance [2].

10.18416/SCP.2025.1936

Electroencephalography (EEG): EEG measures electrical activity in the outer cortical layer (cerebral cortex) of the brain [6]. Our EEG analysis considers four frequency bands: alpha, beta, theta, and gamma [6]. Changes in these frequency bands can reflect cognitive load: e.g., a simultaneous increase in theta activity and a lowered alpha activity indicates an elevated cognitive load [2].

Eye-tracking (EYE): Pupil dilation can serve as an indicator of cognitive load [7]. Larger pupil sizes are associated with higher cognitive load [7]. Studies have shown that pupil size increases with higher cognitive load [8]. The frequency of pupil diameter oscillation can be captured with the index of pupillary activity (IPA) [9].

Each sensor modality offers distinct insights into physiological responses to cognitive load, enabling a more precise and reliable assessment when combined. Consequently, our study aims to evaluate the cognitive load of software developers during realistic, day-to-day tasks by utilizing metrics derived from multi-modal sensor data. Additionally, we strive to identify which physiological metric reflects the subjectively experienced cognitive load.

I.I. Related Work

Physiological multi-modal measurements from software developers have previously been used to evaluate concepts related to cognitive load, such as task difficulty. Using data from wearables, such as EEG, eye tracking, and EDA, and utilizing simple machine learning models like Naïve Bayes, Fritz et al. [2] predicted the perceived task difficulty based on code comprehension tasks completed by developers.

In another study, Fritz and Müller [3] collected data including EDA, eye tracking, EEG, and heart- and breathing-related metrics. To analyze this data, they used machine learning models such as Naïve Bayes and decision trees. They aimed to determine the emotional and cognitive state, their influence on developers' productivity, and the interruptibility of developers. The studies have not directly detected cognitive load by calculating metrics from the sensor data but instead used machine learning [2, 3].

II. Methods and Materials

Dataset: The dataset used in this analysis consists of four software developers performing various tasks reflecting typical daily programming activities to examine cognitive load. All participants were male, right-handed, and aged between 23 and 29, with an average of 25 years. They predominantly programmed in Python, SQL, and C++ and had an average of 5.75 years of programming experience. The experiment comprised four different 10-minute everyday tasks performed in a randomized order. The tasks involved writing Java code, debugging, code documentation, and writing emails. After each task, the participants completed the NASA task load questionnaire (NASA-TLX) [10] to assess their perceived cognitive load. The NASA-TLX Score consists of six scales, comprising the level of frustration, mental demand, effort, physical demand, temporal demand, and performance. The NASA-TLX score was calculated by averaging the six scales [10]. In addition to the tasks, three baseline recordings were conducted at the beginning and end of the study, including participants watching a relaxing fish tank video, a recording with open eyes, and a recording with closed eyes. The experiment workflow was deployed in the JetBrains IDE IntelliJ IDEA to simulate a typical work environment. Participants wore a Shimmer3 GSR+ device to measure EDA and an Emotiv Epoc X (10-20 system) to collect EEG. The Tobii Pro Spark eye tracker was installed below the computer monitor to track the participant's pupil size and gaze. The plugin CognitIDE [11] was employed, which helps to minimize interruptions, ensures smooth recordings during the tasks, and enables a mapping of the physiological activity onto source code. This plugin, written in Kotlin for JetBrains IDEs, allows sensor synchronizations and study flow automation [11]. For planned tasks in the workflow, the plugin automatically starts and stops the recording with sensors synchronized.

Data Analysis: The data underwent a series of preprocessing steps, described in the following sections, to guarantee high-quality data for subsequent interpretation.

EEG: First, the sensor's built-in software removed power line noise [12]. The EEG device also provides a quality evaluation signal for the EEG data ranging from 0 to 4, with four representing the highest quality. We discarded data points with a quality value below four to maintain the integrity and reliability of the analysis, as lower-quality data could introduce noise and inaccuracies into the results (valid data: ≈ 92%). Utilizing the MNE library, the signal was then examined for bad channels, which were treated using spherical spline interpolation [12]. The signal was filtered with a FIR Bandpass filter [13] with a cutoff frequency $[1,30]$ Hz. The filtered signal was re-referenced to its mean and inspected for outliers [12].After the signal preprocessing, the Event-related (De)—/Synchronisation (ERDS) for the alpha and theta band power was calculated to quantify a participant's cognitive load per task (Eq. 1) [14]:

$$\mathrm{ERDS\%} = \frac{\mathrm{BBP} - \mathrm{TBP}}{\mathrm{BBP}} \tag{1}$$

The BBP is the band power of the baseline recording, and the TBP is the band power during a task. A negative ERDS indicates an increase in the band power, whereas a positive ERDS implies a decrease in the band power [14].

EDA: To clean the EDA signal, a fourth-order Butterworth low pass filter with a cutoff frequency of 3 Hz is

10.18416/SCP.2025.1936

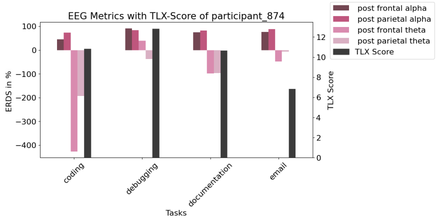

Figure 1: ERDS in % showing the increase and decrease of mean theta and alpha frequency bands of the parietal and frontal brain regions.

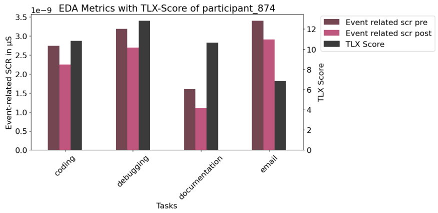

Figure 2: Mean event-related SCR (mESCR) relative to its pre- and post-baseline for all tasks.

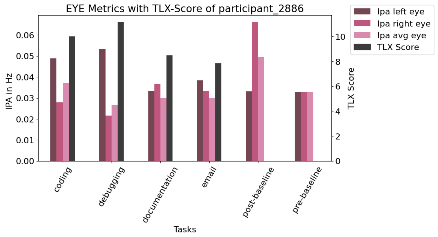

Figure 3: IPA for the left and right eye and the average of both eyes for all tasks and the baseline.

applied [15]. The EDA was then decomposed into its tonic (SCL) and phasic (SCR) components [15] using Neurokit2.

EYE: Blinking or the tracker losing the pupil results in invalid data that must be removed (valid data: ≈ 89%)[16, 17]. The signal was reviewed for outliers, including dilation speed outliers, trend line deviation outliers, temporally isolated samples, and invalid pupil sizes [16]. Identified outliers were removed. Lastly, the baseline correction was executed by subtracting the mean of the cleaned baseline period from the signal. The Index of Pupillary Activity (IPA) was used to quantify the relation between cognitive load and pupil dilation [18]. The metric was computed using the modulus maxima to recognize peaks by applying a discrete mother wavelet transformation.

III. Results and Discussion

As visualized in the figures 1, 2 and 3, the results display indications of cognitive load during the tasks compared to baseline recordings and its alignment with the mean NASA-TLX Score for an exemplary participant.

EEG: Figure 1 shows the ERDS, which visualizes the cognitive load changes in the frontal and parietal lobes during the different tasks based on the mean theta and alpha frequency bands relative to the post-baseline recording. The presented ERDS of the mean alpha band power is strictly positive, whereas the ERDS of the mean theta band power is strictly negative. This indicates that the mean alpha band power desynchronizes and the mean theta band power synchronizes, reflecting changes in cognitive load compared to the baseline [14]. Typically, as cognitive load increases, alpha activity decreases, leading to a rise in alpha's ERDS, while theta activity increases, resulting in a decrease in theta's ERDS [14]. Higher theta activity and a negative ERDS indicate a high cognitive load. The coding, documentation, and email tasks show the expected relationship with cognitive load, especially in theta activity. However, the debugging task corresponds to the highest NASA-TLX score and points towards the lowest theta wave activity. A possible reason for this unexpected result is cognitive overload. The debugging task may have been too difficult, leading them to give up, resulting in low measured cognitive load. Additionally, numerous participants indicated a lack of familiarity with the IDE environment, which heightened the difficulty, as they mainly did not program in Java, and thus did not know how to use the IDE's debugger effectively. The findings express the ERDS is a reliable metric to detect an increased cognitive load relative to a baseline, however it fails to detect cognitive overload.

EDA: Figure 2 shows the mean event-related SCR (mESCR), which increased during tasks compared to pre- and post-task baseline recordings. Higher mESCR values suggest increased cognitive load [2], which are consistent with participants' NASA-TLX ratings, indicating the metric's reliability.The SCR did not always rise with the NASA-TLX scores, possibly due to signal quality issues. The Shimmer3 GSR+ electrodes worn around the finger may have recorded distorted values whilst typing on the keyboard during the tasks.

EYE: As visualized in Figure 3, the results showed an expected relation between the IPA of the left eye and the NASA TLX score. However, the highest NASA TLX score corresponds to the lowest IPA of the right eye, whereas the second lowest score (documentation task) resembles the highest IPA of the right eye. In general, higher IPA correlates with a higher cognitive load [18]. The differences in the IPA of the eyes are relatively small for most tasks, except for the debugging task. Considering the differences, the IPA might not be best metric for indicating the cognitive load. Future research should investigate the differences in the IPA of the eyes while doing complex tasks such as debugging. A study [9] discovered that IPA does not correlate with task difficulty, referring to the fact that IPA values may not be reliable for complex tasks.

The metrics calculated from sensor data, recorded and synchronized using CognitIDE, indicated changes in cognitive load that align with participants' subjective perceptions during simulated everyday working tasks.

10.18416/SCP.2025.1936

IV. Conclusion

In conclusion, this exploratory study analyzed the relationship between physiological signals and the perceived cognitive load of software developers using the metrics IPA, mESCR, and ERDS calculated from the pupil dilation, EDA and EEG signals. Based on the above analyses, mESCR best reflects the cognitive load. The results indicate that cognitive load is partially reflected in physiological data collected from wearables. This suggests the potential of wearables to enhance cognitive load recognition, serving as an extension to traditional questionnaires. Validation with a larger sample size and over an extended period in real-life working conditions is necessary to ensure the credibility of the results. In a future work, the relationship between cognitive load and physiological metrics will analyzed more closely with a larger sample size and by utilizing machine learning approaches. This is a next step towards finding measures to support software developers during their cognitively demanding working tasks.

Acknowledgments

The work was carried out at the Hasso Plattner Institute, Connected Healthcare in cooperation with Prof. Dr.-Ing. Marcin Grzegorzek, Institute of Medical Informatics, Universität zu Lübeck. Additional thanks to Md Abdullah for his resources. Research funding: Charlotte Brandebusemeyer's work is funded by the SAP - HPI Research Program, and Fabian Stolp's work is funded by the Hasso Plattner Institute Research School on Data Science and Engineering.

Author's statement

Authors state no conflict of interest. Informed consent to a voluntary participation has been obtained from all individuals included in this study. The research related to human use complies with all the relevant national regulations and institutional policies and was performed in accordance with the tenets of the Helsinki Declaration and has been approved via an ethics approval by the ethics board of the University of Potsdam. DeepL and Grammarly gave linguistic support.

References

[1] P. Chandler and J. Sweller. Cognitive load theory and the format of instruction. *Cognition and Instruction*, 8(4):293–332, 1991, doi:10.1207/s1532690xci0804_2.

[2] T. Fritz, A. Begel, S. C. Müller, S. Yigit-Elliott, and M. Züger. Using psycho-physiological measures to assess task difficulty in software development. *Proceedings of the 36th International Conference on Software Engineering*, 2014, doi:10.1145/2568225.2568266.

[3] T. Fritz and S. C. Müller, Leveraging biometric data to boost software developer productivity, in *2016 IEEE 23rd International Conference on Software Analysis, Evolution, and Reengineering (SANER)*, 5, 66–77, 2016. doi:10.1109/SANER.2016.107.

[4] H. Critchley and Y. Nagai, Electrodermal activity (eda), in *Encyclopedia of Behavioral Medicine*. Springer New York, 2013, 666–669. doi:10.1007/978-1-4419-1005-9_13.

[5] M. Benedek and C. Kaernbach. A continuous measure of phasic electrodermal activity. *Journal of Neuroscience Methods*, 190(1):80–91, 2010, doi:10.1016/j.jneumeth.2010.04.028.

[6] H. Kessler, Das Elektroenzephalogramm (EEG), in *Kurzlehrbuch Medizinische Psychologie und Soziologie*, 4., überarbeitete Auflage, Georg Thieme Verlag KG, 2021. doi:10.1055/b-003-104356.

[7] E. Granholm, R. F. Asarnow, A. J. Sarkin, and K. L. Dykes. Pupillary responses index cognitive resource limitations. *Psychophysiology*, 33(4):457–461, 1996, doi:10.1111/j.1469-8986.1996.tb01071.x.

[8] S. Moresi, J. J. Adam, J. Rijcken, P. W. Van Gerven, H. Kuipers, and J. Jolles. Pupil dilation in response preparation. *International Journal of Psychophysiology*, 67(2):124–130, 2008, doi:10.1016/j.ijpsycho.2007.10.011.

[9] P. Weber, F. Rupprecht, S. Wiesen, B. Hamann, and A. Ebert, Assessing cognitive load via pupillometry, in *Advances in Artificial Intelligence and Applied Cognitive Computing*, 1087–1096, Cham: Springer International Publishing, 2021. doi:10.1007/978-3-030-70296-0_86.

[10] S. G. Hart and L. E. Staveland, Development of nasa-tlx (task load index): Results of empirical and theoretical research, in *Human Mental Workload*, ser. Advances in Psychology, P. A. Hancock and N. Meshkati, Eds., 52, North-Holland, 1988, 139–183. doi:10.1016/S0166-4115(08)62386-9.

[11] F. Stolp, M. Stellmacher, and B. Arnrich, Cognitide: An IDE plugin for mapping physiological measurements to source code, in *Companion Proceedings of the 32nd ACM International Conference on the Foundations of Software Engineering, FSE 2024, Porto de Galinhas, Brazil, July 15-19, 2024*, 592–596, ACM, 2024. doi:10.1145/3663529.3663805.

[12] N. B. Shamlo, T. R. Mullen, C. Kothe, K. Su, and K. A. Robbins. The prep pipeline: Standardized preprocessing for large-scale eeg analysis. *Frontiers in Neuroinformatics*, 9, 2015, doi:10.3389/fninf.2015.00016.

[13] J. X. Suárez-Revelo, J. F. Ochoa-Gómez, and C. A. Tobón-Quintero, Validation of EEG pre-processing pipeline by test-retest reliability, in *Applied Computer Sciences in Engineering - 5th Workshop on Engineering Applications, WEA 2018, Medellín, Colombia, October 17-19, 2018, Proceedings, Part II*, 916, 290–299, Springer, 2018. doi:10.1007/978-3-030-00353-1_26.

[14] P. Antonenko, F. Paas, R. Grabner, and T. van Gog. Using electroencephalography to measure cognitive load. *Educational Psychology Review*, 22(4):425–438, 2010, doi:10.1007/s10648-010-9130-y.

[15] S. Subramanian, R. Barbieri, and E. N. Brown. A systematic method for preprocessing and analyzing electrodermal activity. *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 6902–6905, 2019, doi:10.1109/EMBC.2019.8857757.

[16] S. Mathôt, J. Fabius, E. Van Heusden, and S. Van der Stigchel. Safe and sensible preprocessing and baseline correction of pupil-size data. *Behavior Research Methods*, 50(1):94–106, 2018, doi:10.3758/s13428-017-1007-2.

[17] M. E. Kret and E. E. Sjak-Shie. Preprocessing pupil size data: Guidelines and code. *Behavior Research Methods*, 51:1336–1342, 2018, doi:10.3758/s13428-018-1075-y.

[18] A. T. Duchowski, K. Krejtz, I. Krejtz, C. Biele, A. Niedzielska, P. Kiefer, M. Raubal, and I. Giannopoulos, The index of pupillary activity: Measuring cognitive load vis-à-vis task difficulty with pupil oscillation, in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–13, 2018. doi:10.1145/3173574.3173856.