Computers in Human Behavior Reports 4 (2021) 100116

Contents lists available at ScienceDirect

Computers in Human Behavior Reports

journal homepage: www.journals.elsevier.com/computers-in-human-behavior-reports

Robust cognitive load detection from wrist-band sensors

Vadim Borisov *, Enkelejda Kasneci, Gjergji Kasneci

The University of Tübingen, Germany

ARTICLE INFO

Keywords:
Cognitive load detection
Human-computer interaction
Machine learning
Ensemble methods

ABSTRACT

In recent years, the detection of cognitive load has received a lot of attention. Understanding the circumstances in which cognitive load occurs and reliably predicting such occurrences, offers the potential for considerable advances in the field of Human-Computer Interaction (HCI). Numerous HCI applications, ranging from medical and health-related solutions to (smart) automotive environments, would directly benefit from the reliable detection of cognitive load. However, this task still remains highly challenging. We present a machine learning (ML) approach based on ensemble learning for robust cognitive load classification. The features used by the proposed solution are generated from the interpretation of physiological measurements (e.g., heart rate, r-r interval, skin temperature, and skin response) from a wearable device. Hence, our approach consists of two steps: (1) transforming the original data into discriminative features and (2) training an ensemble model to accurately and robustly predict cognitive load. The empirical results confirm that our method has a superior performance compared to various state-of-the-art baselines on the original and transformed data. Moreover, in the open-data CogLoad@UbiComp 2020 Competition, the proposed approach achieved the best results among 17 competing approaches and outperformed all participating competitors by a considerable margin.

1. Introduction

The degree to which our cognitive resources, e.g., attention, working memory, decision making, or task-related knowledge, are currently used is commonly referred to as cognitive load (Sweller, 2011). With emerging novel multimedia technologies, cognition-aware computing, and human-centered systems that aim to automatically adapt to the user's cognitive state, predicting and quantifying cognitive load can be beneficial in various applications of Human-Computer Interaction (HCI). More specifically, the ability to estimate the proximal zone of a user, where stress, frustration (as typically resulting from cognitive overload), and boredom (originating from low levels of cognitive load) can be avoided, promises to open new avenues towards the development of truly intelligent user-centered systems and enhanced user experience. For example, intelligent tutoring systems could have crucially aided learning and teaching in the current COVID-19 epidemiological situation. Beyond the learning context, in entertainment-related applications such as gaming, the online assessment of a user's cognitive load could contribute significantly to an enhanced user experience. In various medical applications, the ability to detect the cognitive overload of medical experts could help in the development of appropriate supporting measures and systems. Considered a highly important measure towards a better understanding of human cognition and performance, the measurement and prediction of cognitive load has been the focus of research works for more than three decades.

A variety of approaches to cognitive load have been explored during the past decades (Kramer, 1990; Sweller, 2011), ranging from questionnaire-based techniques such as the NASA TLX (Hart & Staveland, 1988) self-report to advanced methods based on Deep Neural Networks on image data (Fridman et al., 2018). The main limitation of self-reports, however, is the subjective nature of the user responses, which hinders the identification of ground truth labels. Furthermore, it has been shown that such questionnaires may induce additional load to the user (Abdelrahman et al., 2017) and are not applicable to online settings where a person's cognitive load has to be estimated during task performance. Therefore, with increasing technological possibilities for user monitoring (e.g., through electroencephalography, eye-tracking technology, camera-based user monitoring, or galvanic skin response and heart rate sensors), estimating cognitive load based on physiological or image data of a user has progressively gained research focus.

This work aims to achieve the automated detection of cognitive load on physiological user data sensed *non-invasively* from a wearable using machine learning techniques. More specifically, we present the 1st place solution from the CogLoad@UbiComp 2020 Competition, which

* Corresponding author.
E-mail addresses: vadim.borisov@uni-tuebingen.de (V. Borisov), enkelejda.kasneci@uni-tuebingen.de (E. Kasneci), gjergji.kasneci@uni-tuebingen.de (G. Kasneci).

https://doi.org/10.1016/j.chbr.2021.100116
Received 10 February 2021; Received in revised form 22 June 2021; Accepted 23 June 2021
Available online xxxx

addressed cognitive load detection from low-cost wrist-band sensors (Microsoft Band 2), i.e. measuring galvanic skin response (GSR), skin temperature (ST), heart rate (HR), and heart rate variability (RR intervals). Our method consists of two steps; first, data is transformed from temporal to static data, allowing all standard ML algorithms to be used. The second step is ensemble learning using decision tree-based models, which is proved to be robust on various tasks, especially on data.

In summary, the contributions of this work are multi-fold:

- We present the winning solution of the CogLoad@UbiComp 2020 Competition. The proposed method outperformed all other state-of-the-art approaches from competitors in terms of predictive performance on unseen test data by a considerable margin.
- We demonstrate how simple yet effective data transformation techniques can improve the state-of-the-art machine learning approaches, thereby advancing the current state-of-the-art in the area of cognitive load detection based on wearable sensors.
- We provide a comprehensive comparison of the proposed ensemble approach with other state-of-the-art machine learning methods on the original and transformed data.
- An open-source implementation of our approach can be found here: https://www.github.com/unnir/CogLoad_UbiComp2020.

The remainder of this article is organized as follows. In Section 2, we first provide an overview of related work on cognitive load detection from physiological data, and especially on methods approaching the CogLoad@UbiComp 2020 Challenge. The cognitive load data set from this challenge is described in detail in Section 3. Section 4 presents our feature transformation methods, followed by our model description for accurate cognitive load detection in Section 5. The experimental evaluation and performance results of our model are presented and discussed in Section 6. In Section 7, we discuss the limitation of our approach with future steps. Section 8 concludes this work.

2. Related work

Beyond questionnaires and self-reports, user monitoring holds enormous potential for the robust measurement of a person's cognitive load in an online fashion, i.e. during task performance. This information can, in turn, be employed by the systems with which the user is interacting and, thus, become beneficial for the user's performance through adaptation or additional supportive measures. Various approaches have therefore been proposed during in recent years to tackle the detection of cognitive load and will be discussed in the following along with specific solutions and implementations on the same data set of the CogLoad@UbiComp2020 Challenge.

2.1. Detecting cognitive load from brain activity

Common neuro-imaging techniques that have been used to detect cognitive load are electroencephalography (EEG), e.g., as in (Friedman et al., 2019; Mills et al., 2017), near-infrared spectroscopy (NIRS, e.g., in (Grubov et al., 2020; Keshmiri et al., 2017)), or functional magnetic resonance imaging (fMRI, as for example from (Mäki-Marttunen et al., 2019)). Although brain imaging techniques promise to deliver highly accurate detection and prediction of cognitive load, the technology has not yet reached the stage of ubiquity and low-cost availability. These techniques are therefore not yet applicable to use-cases outside the laboratory.

2.2. Detecting cognitive load from eye movements and pupil information

With recent advances in eye-tracking technology, a non-intrusive way to infer information about a person's cognitive state is becoming available. In recent years, different features of the eye and of eye movements have been investigated as predictors of cognitive load. For example (Chen et al., 2011; He & McCarley, 2010; Inamdar & Pomplun, 2003; Van Orden et al., 2001; Wang et al., 2014), associated longer fixations with more effort and thus with higher levels of cognitive load. However, other streams of related work have indicated the opposite, i.e., by relating longer fixation to lower cognitive effort, e.g., (Amadieu et al., 2009; Van Gog et al., 2005). These results, however, might be influenced by processing difficult or visually challenging stimuli (Rayner, 1998). In addition, a few recent articles have looked at the relationship between smooth pursuits and cognitive load (Kosch et al., 2018; Stubbs et al., 2018), and reported a high predictive power of features extracted from smooth pursuits on cognitive load. In the past few years, enabled through high-speed eye-tracking devices, a research line in this area has investigated the relationship between microsaccades and cognitive load. Microsaccades describe small, involuntary eye movements which occur during a fixation period and are assumed to be highly associated with cognitive and visual load. More specifically, tasks that induce high visual load were found to cause an increase of the frequency of microsaccades (Benedetto et al., 2011), while auditory or arithmetic tasks have been found to have the opposite effect, i.e., reduce their frequency (Gao et al., 2015; Krejtz et al., 2018; Siegenthaler et al., 2014). In addition to eye-movement characteristics, blinks and their frequency have also been investigated regarding their relation to visual or cognitive load (Bristow et al., 2005; Fukuda et al., 2005). More specifically, it has been reported that higher cognitive load induces more frequent blinking (Chen & Epps, 2014; Hogervorst et al., 2014).

Another line of research in this area focuses on predictive features of cognitive load derived from the eye pupil and its oscillations. It is well understood that high cognitive load causes characteristic patterns of pupil dilation (Beatty et al., 2000; Kramer, 1991), which is also known as the task-evoked pupillary response. It has been further shown that this effect even persists within a task, between tasks, and between individuals, concluding that there is a consistent influence of cognitive load on the pupil diameter (Kahneman, 1973). Since changing illumination conditions (e.g., environmental lightning or lightning changes of the visual stimulus itself) also affect pupillary response, several research articles have focused on generalizable approaches, e.g., (Appel et al., 2018, 2019; Duchowski et al., 2018, pp. 1–13; Faure et al., 2016; Kun et al., 2013; Marshall, 2000, 2007; Palinko et al., 2010; Pfleging et al., 2016).

2.3. Image-based approaches to cognitive load detection

Recent approaches have proposed (deep-learning) frameworks for cognitive load detection based on image data (Fridman et al., 2018; Rafiqi et al., 2015) or thermal images (Abdelrahman et al., 2017). The latter group of methods aims to automatically detect changes in skin temperature, respiration, or heart rate related to changing levels of cognitive load based on thermal images. Although image-based technology is, in general (primarily due to its non-intrusive nature), highly interesting and relevant, further research is required to determine the generalizability of such approaches across tasks and subjects.

2.4. Physiological signals for cognitive load estimation: galvanic skin response (GSR), heart rate (HR), and heart rate variability (HRV)

Physiological sensors measuring GSR, HR, and HRV[1] are meanwhile available at low cost and large scale. Thus, these signals can be employed for user monitoring in a variety of applications. As a measure of skin conductivity, GSR is considered a strong indicator of stress and cognitive load and, as such, is used in various approaches to detect cognitive load. More specifically, various studies have reported an increase in GSR with

[1] Note that HRV corresponds to R-R intervals. We use the short form RR in the following sections, which is in accordance with the terminology used in the data set description of the CogLoad@UbiComp 2020 Challenge.

increasing cognitive load, e.g. (Kasneci et al., 2017; Nourbakhsh et al., 2012, 2017; Shi et al., 2007). Cardiac measures such as HR and HRV have also been successfully employed in many related studies on cognitive load detection, e.g. (Gjoreski et al., 2018; Hughes et al., 2019; Kübler et al., 2014; McDuff et al., 2014; Mehler et al., 2011; Wang & Guo, 2019).

Apart from the above-mentioned lines of related research, various approaches have proposed multimodal methods of cognitive load detection to increase the accuracy of predictions (Debie et al., 2021). Prabhakar et al. (2020) suggested the estimation of cognitive load from eye-movement and pupil dilation parameters. In four independent studies with 123 participants, Sharma et al. (2020) analysed the assessment of cognitive load through physiological responses and facial expressions, a method that works well for recognizing successful perception, at least by aggregating physiological and eye-tracking signals (Kasneci et al., 2017). Finally, in the aviation industry, the estimation of the cognitive load plays a crucial role and has therefore been analysed in various studies, e.g., (Babu et al., 2019; Di Nocera et al., 2007; Wilson et al., 1994).

Our work specifically addresses the applicability of machine learning models to the prediction of the cognitive load of a user from physiological data collected from smart wearables, since these sensors are not only non-intrusive and affordable, but also convenient to use. Hence, in the following, we will briefly discuss the state-of-the-art in this area as well as competing approaches from the same challenge.

2.5. Cognitive load estimation from wearable sensors

Setz et al. (2010) introduced a machine learning approach in 2009 based on six classifiers to distinguish stress from cognitive load in an office environment based on data from 33 subjects in a laboratory intervention study. Their methods achieved an accuracy of 82.8%, thus exhibiting promising steps towards cognitive load detection based on low-cost smart devices. Huang et al. (2018) showed that physiological signals from wearable devices could be used to analyze psychological factors that can help with disease prevention. In another study, Schaule et al. (2018) introduced a system coined COLLINS (COgnitive Load CLassification to prevent INterruptionS), which utilized a smart wrist-band device for cognitive load estimation. This approach showed that the cognitive load could be estimated using sensors from a smart device. In an evaluation with ten subjects the authors compared three machine learning algorithms - SVM, Random Forest, and Naive Bayes (SVM and Random Forest are selected as baselines for our experiments) and reported an accuracy between 66% and 86% for individual participants. For the general classification task, COLLINS achieved an accuracy between 32% and 36% in a ten-fold cross-validation. The work from (Gjoreski et al., 2018), based on 25 volunteers, utilized sensor information from a simple wearable device in order to measure the cognitive load. In parallel, they collected physiological data with a device, extracted features, and then constructed machine learning models for cognitive load prediction. Although the final accuracy of the statistical model was only 51%, the work confirmed that it is possible to estimate cognitive load using a wearable device. The data collected by the authors was later released to the research community as a challenge and served as a data foundation of our work.

2.6. Related approaches from the CogLoad@UbiComp 2020 challenge

The CogLoad@UbiComp 2020 Competition, along with an open data set, was advertised by the 5th International Workshop on Smart & Ambient Notification and Attention Management (UbiTention 2020) at UbiComp 2020 (Li & De Cock, 2020) (team Lynx from the University of Washington). used different combinations of feature engineering steps (e.g., Fast Fourier Transforms, Sliding Mean Filter) in combination with machine learning models (e.g., Logistic Regression, (Boosted) Decision Trees, Random Forests, and Support Vector Machines) to approach the problem. Their best processing pipelines yielded an accuracy of 63% on the data set, which is in line with previous work on smartwatch data (Gjoreski et al., 2020). In another work on this data set, Salfiger (Salfinger, 2020) investigated the applicability of deep learning approaches for cognitive load monitoring. More specifically, the author evaluated different configurations of Recurrent Neural Networks (Schmidhuber, 2015), namely Gated Recurrent Units (GRUs) and RNNs using Long short-term memory (LSTM) cells, and found that architectures based on GRUs achieve the best performance. A limitation of this approach is however the complexity of the models which have a tendency to overfit with respect to the small size of the data set from the challenge, which, in addition, is characterized by considerable between-subject variance and subject-related bias (Salfinger, 2020). The team HCM-lab from the University of Augsburg, Germany, got second place in the competition by using a deep learning-based approach. First, they trained an autoencoder on original data. Second, they utilized the encoder only as an input block and used a three-layer artificial neural network on top of it. A team from the VTT Technical Research Centre of Finland (Tervonen et al., 2021) used a support-vector machine (SVM) based (Boser et al., 1992) approach with the Bayesian optimization step. Their best mode received the 3rd place in the CogLoad@UbiComp 2020 competition.

3. The data set

The method we propose in this work was developed and evaluated in the context of the CogLoad challenge (van Berkel et al.) using the CogLoad data set (Gjoreski et al., 2020) provided by UbiComp 2020 (international joint conference on pervasive and ubiquitous computing, 2020), the leading venue in the area of ubiquitous and pervasive computing. The data set is freely available online.[2]

The data set contains four different physiological measurements, which were recorded by a Microsoft Band 2 wrist-band from 23 participants performing six psychological tasks on a PC with varying levels of difficulty, as well as measurements recorded while the participants were in a rest state. Participants' mean age was 29.51 (standard deviation is 10.10). The right was the dominant hand of 22 participants, while 1 participant was left-handed. All participants had the wrist-band device strapped to their left hand. In the conducted trials, the participants solved cognitive tasks of varying difficulty. The experiments were conducted in a quiet and normal-temperature room with one participant at a time. The experimental scenario consisted of two parts. Part 1 was devoted to estimating the participants' cognitive capacity. For assessing the participants' cognitive capacity, the participants solved two N-back tasks (Schmiedek et al., 2014), i.e., 2-back and 3-back tasks, with a 3-min rest after each of them. In Part 2, the participants were presented with six primary tasks. For each task, three variations of a randomly selected primary cognitive-load task were presented to the participant. The variations differed in complexity (easy, medium, and difficult). More information on participants and task are provided in (Gjoreski et al., 2020).

For each participant, 50% of the samples correspond to the cognitive load state, the other 50% to resting. In the data set, the target variable is represented by a binary value, i.e., a ‘1’ represents a low cognitive load, i.e., resting, and a ‘0’ represents a high cognitive load.

The physiological measurements include Galvanic skin response (GSR), heart rate (HR), R-R intervals (RR), and skin temperature (ST). All these measurements were sampled at a sampling rate of 1Hz. From the 23 participants, the recorded measurements of 5 participants were used for testing and the measurements of the remaining 18 participants for training. The training and test data were generated using time windows of 30 s.

The sensor files in the training data (GSR, HR, RR, and ST) contain 632 lines x 30 columns, corresponding to 632 instances each containing 30 samples (i.e., generated within 30 s at the sampling rate 1Hz). The training instances are randomly permuted. The sensor files in the test

[2] CogLoad@UbiComp Data set: https://www.ubittention.org/2020/.

data (GSR, HR, RR, and ST) contain 193 lines x 30 columns, corresponding to 193 instances each containing 30 samples (30 s at the sampling rate 1 Hz). The test instances are also randomly permuted.

Note, that apart from sampling and resampling, no additional preprocessing steps (e.g., a fast Fourier transform (FFT) filtering) were used. Thus, the data is raw as provided by the Microsoft Band 2. As an example, the data for one of the participants is visualized in Fig. 1.

4. Transformation of the raw data into discriminative features

Each instance from the raw training and test data set can be represented as a vector of length 120, i.e., as a concatenation of the vectors of length 30 from the single physiological signals recorded in the 30-s window. More specifically, let $\mathcal{I}$ denote all instances in the training and test set. For each instance $i \in \mathcal{I}$, we define a vector $\mathbf{x}_i = (\mathbf{x}_{\mathbf{GSR}i}, \mathbf{x}_{\mathbf{HR}i}, \mathbf{x}_{\mathbf{RR}i}, \mathbf{x}_{\mathbf{ST}i}) \in \Re^{120}$, where $\mathbf{x}_{\mathbf{GSR}i}, \mathbf{x}_{\mathbf{HR}i}, \mathbf{x}_{\mathbf{RR}i}, \mathbf{x}_{\mathbf{ST}i} \in \Re^{30}$, represent the raw vectors from the measurements of GSR, HR, RR, and ST, respectively, within the 30-s window for the same instance *i*. We denote the set of all these raw vectors by $\mathcal{X}_{\mathcal{I}} = \{\mathbf{x}_i | i \in \mathcal{I}\} \subset \Re^{120}$.

Certainly, the raw vectors generated in this way can readily be used in combination with state-of-the-art classification techniques with the goal of recognizing cognitive load (see also Section 6). However, given their high dimensionality, their time-series character, and the relatively small size of the training data set, i.e., low number of instances, it is often practical to transform the raw feature vectors into lower-dimensional feature vectors that carry the majority of important information from the raw data. This way, we can avoid the dimensionality problem, also known as the Curse of Dimensionality (Verleysen & François, 2005), and can develop highly discriminative and robust classifiers. To this end, we are interested in a transformation function $\Phi_{\mathcal{F}} : \mathcal{X} \rightarrow \Re^k, k \ll 120$, such that $\mathcal{F} = \{f_{GSR1}, f_{HR1}, f_{RR1}, f_{ST1}, \ldots, f_{GSRk}, f_{HRk}, f_{RRk}, f_{STk}\}$ is a family of feature functions of the form $f_{arg} : \Re^{30} \rightarrow \Re$, and

$$\begin{aligned}\Phi_{\mathcal{F}}(\mathbf{x}) &= (f_{GSR1}(\mathbf{x}_{\mathbf{GSR}}), f_{HR1}(\mathbf{x}_{\mathbf{HR}}), f_{RR1}(\mathbf{x}_{\mathbf{RR}}), f_{ST1}(\mathbf{x}_{\mathbf{ST}}), \ldots,\\ &\quad f_{GSRk}(\mathbf{x}_{\mathbf{GSR}}), f_{HRk}(\mathbf{x}_{\mathbf{HR}}), f_{RRk}(\mathbf{x}_{\mathbf{RR}}), f_{STk}(\mathbf{x}_{\mathbf{ST}}))\end{aligned}$$

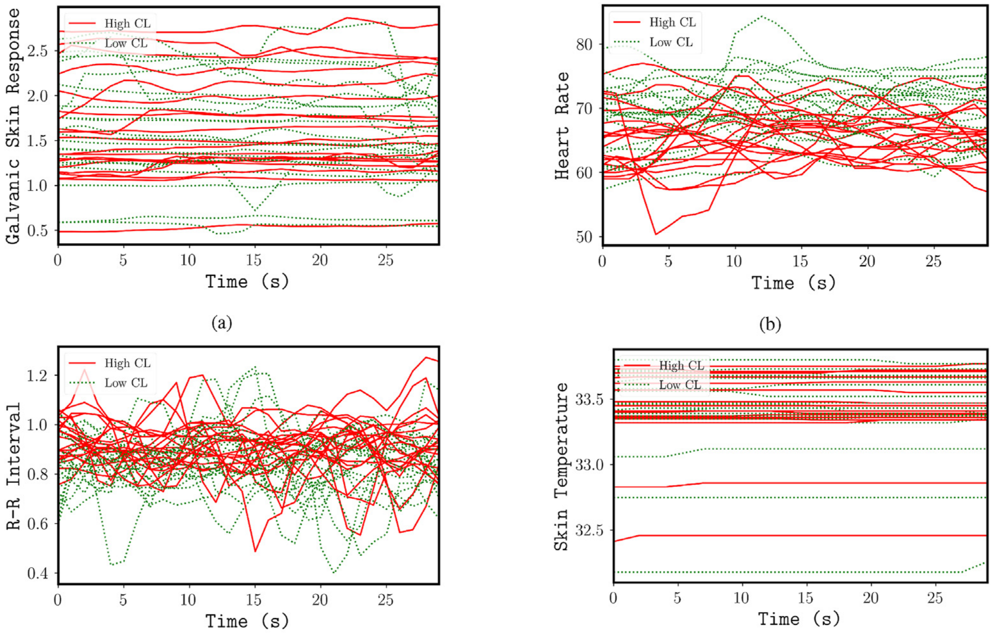
There exist various possibilities to construct an adequate family of feature functions $\mathcal{F}$ for the described data set. Various approaches ranging from Wavelet and Fourier Transformations (Bloomfield, 2004; Chan & Fu, 1999; Chaovalit et al., 2011; Grinsted et al., 2004) to grammar-based evolutionary approaches (De Silva & Leong,) are able to deal with the time-series character of the data and generate low-dimensional feature vectors. For an overview, we refer the reader to (Fu, 2011). In fact, at least one of the competitors (Li & De Cock, 2020) in the CogLoad@UbiComp 2020 Competition builds on some of the mentioned feature generation strategies for the competing solution.

However, the relatively small size of the training data in the CogLoad@UbiComp 2020 Competition poses a serious challenge and does not allow these strategies to generate discriminatory features for robust predictions. Hence, after some analysis on feature generation strategies, we decided to use simple aggregation statistics as feature functions for $\mathcal{F}$. These include the *minimum*, the *maximum*, the *mean*, the *median*, the *standard deviation*, the *sum*, and the *skew* values from the 30-s sequence, for each of the 4 physiological signals (i.e., for GSR, HR, RR, and ST), respectively. The feature generation and model development pipeline for the solution presented in this article is depicted in Fig. 2.

Finally, we employed the CancelOut mechanism (Borisov et al., 2019) and the Shapley additive explanations (SHAP) framework (Lundberg & Lee, 2017) for the analysis of feature importance. The results of this analysis are shown in Fig. 5 and indicate that the vast majority of the proposed features (generated from the raw data) contain highly discriminative information for the prediction task at hand. As expected, the standard deviation of the physiological signals consistently appears among the most discriminative features.

5. A robust predictive method for cognitive load detection

In this section, we present our ensemble learning approach to robust and accurate cognitive load detection based on the CogLoad@UbiComp 2020 data set. The pipeline of feature generation and ensemble model development is summarized in Fig. 2.



(a)

(b)

(c)

(d)

Fig. 1. Data set visualization for a single participant; the subfigures show galvanic skin responses (a), heart rates (b), R-R Intervals (c), and skin temperature (d). Red lines indicate a high cognitive load, and green lines represent a resting state, respectively. From the visual inspection, it can be seen that there are no clear distinctions between the two states. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

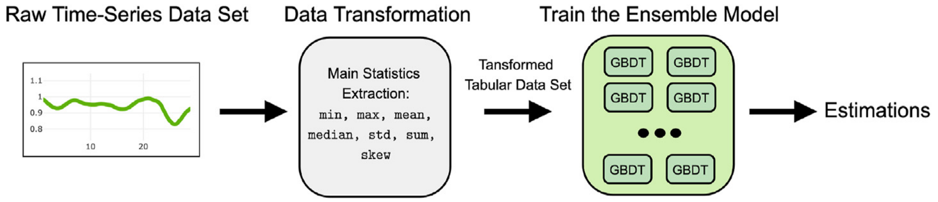


Fig. 2. Pipeline of feature generation and ensemble model development.

5.1. Base learners

After conducting an empirical evaluation of various machine algorithms as base learners for our ensemble approach, we selected the Gradient Boosting Decision Trees (GBDT) algorithm (Friedman, 2002). Our empirical findings on the excellent predictive performance of GBDT are also supported and complemented by previous results from numerous Data Science competitions and challenges. According to (Chen & Guestrin, 2016), in 2015, among the 29 winning solutions of Kaggle challenges (Kaggle.com), 17 solutions built on the GBDT algorithm.

5.2. Ensemble model

For the proposed ensemble solution, we used eight GBDT models. We built each of the eight models based on the LightGBM implementation (Ke et al., 2017). Prior studies have shown that the diversity of base learners clearly helps to reduce bias (which is very important for small data sets) and improve the overall performance of the ensemble algorithm (Kuncheva & Whitaker, 2003; Rokach, 2010). Thus, we trained the base learners by using different hyper-parameters. For the hyper-parameters selection, we utilized the random search and Bayesian optimization strategies (Bergstra et al., 2013). Since adding more models to the ensemble did not improve the predictive performance, the final meta-model consists of eight different GBDT models, where the final prediction is the mean value from all these models.

5.3. Validation

To provide robust estimations and exploit the training data as effectively as possible, we adopted an out-of-fold (OOF) cross-validation strategy. More specifically, from the folds that are used for validation during the cross-validation, we randomly generated hold-out samples, which served as unseen test examples. Based on these hold-out samples, we can estimate the predictive performance of each of the eight GBDT models on unseen data. For the cross-validation we employed a two iteration of stratified 5-fold cross-validation.

6. Experimental evaluation

In this section, we provide an overview of the experimental setup. More specifically, we describe the state-of-the-art predictive approaches that we have considered in the evaluation as baseline models and present the results of their predictive performance on the original feature vectors from $\mathcal{X}$, as well as on the transformed feature vectors using $\Phi_{\mathcal{F}}$ (as described in Section 4), in comparison to the approach proposed in this article.

6.1. State-of-the-art predictive algorithms as baseline approaches

In our experiments, we compare the proposed ensemble model with following predictive algorithms:

- Logistic Regression (LR) (Friedman et al., 2001). LR is a linear classifier with surprisingly strong predictive performance on many practical use cases. Moreover, the parameters of LR can also be fitted quite well on small-size training data.
- k-nearest neighbors (kNN) (Friedman et al., 2001). The kNN classifier often shows a nice predictive performance in practice and comes with strong theoretical guarantees on the possible classification error (more specifically, the Bayes error rate (Hastie et al., 2009)).
- Support-Vector-Machine (SVM) (Boser et al., 1992). The SVM classifier belongs to the most rigorously analysed and refined algorithms (e.g. (Hastie et al., 2009; Scholkopf & Smola, 2018),) and is often one of the best-performing predictive methods in practice.
- Random Forest (RF) (Breiman, 2001). The RF algorithm constructs an ensemble of decision trees that are sufficiently different from each other, allowing the RF to achieve a significantly higher predictive performance than individual decision trees (Hastie et al., 2009).
- Adaptive Boosting (AdaBoost) (Freund & Schapire, 1997). The AdaBoost algorithm is a highly versatile approach and, despite its simplicity, it works astonishingly well in practice, ranking it among the best performing ML methods.
- Gradient-Based Decision Trees (GBDT) (Friedman, 2002). The GBDT algorithm has proven to be one of the best performing predictive methods on heterogeneous tabular data (Chen & Guestrin, 2016). Its generalization and practical capabilities to handle missing values as well as variance and bias in the feature values make it one of the most valuable machine learning algorithms.
- Multilayer Perception (MLP) (Gardner & Dorling, 1998). Given the current popularity of neural networks, despite the relatively small data set, we felt that a corresponding approach should definitely be considered as a baseline. To this end, we decided to use a two-layer, fully-connected artificial neural network, the weights of which are optimized through stochastic gradient descent.

For all above algorithms, except GBDT, we used the scikit-learn library (Pedregosa et al., 2011). For GBDT, we selected the LightGBM implementation (Ke et al., 2017). For the hyper-parameters selection, we utilized the random search and Bayesian optimization strategies (Bergstra et al., 2013).

6.2. Evaluation measures

Following the instructions of the competition, two evaluation measures were used for the experiments:

- Accuracy, which is defined as $Accuracy = \frac{TP+TN}{TP+FN+TN+FP}$, where TP and TN are the true positive and true negative numbers, respectively. FP and FN represent the numbers of misclassifications for the negative and the positive class, respectively. Accuracy measures the fraction of correct predictions and works well for data sets in which the class frequencies are balanced, which is the case for the CogLoad@UbiComp 2020 data set. However, for imbalanced data sets, accuracy can be quite misleading.
- ROC-AUC (i.e., the area under the receiver operating characteristic curve), which quantifies the performance of a classification model over all classification score thresholds. The ROC curve plots two parameters: (1) the True Positive Rate, i.e., $tpr = \frac{TP}{TP+FN}$, and (2) the False Positive Rate, i.e., $fpr = \frac{FP}{FP+TN}$. Note that the tpr is a synonym for the recall of a predictive algorithm, whereas the fpr represents the rate of false alarms. The ROC curve plots the tpr vs. the fpr values at different

classification score thresholds. It can be shown that the area under the ROC curve is the ranking accuracy with respect to the classification score returned by a classifier. Ideally, instances that belong to the positive class should be assigned a higher score by the classifier and thus ranked higher than the instances that belong to the negative class. Hence, AUC of 1 means that all positive instances are ranked before the negative instances and the two classes are clearly separated by the classifier.

6.3. Evaluation results

In order to evaluate both the proposed ensemble method and the features generated by the proposed data transformation scheme, we conducted experiments using two data sets. The first data set consists of the original feature vectors, which are based on the raw (time-series). The second data set consists of the transformed feature vectors, i.e., using the $\Phi_{\mathcal{F}}$-transformation as described in Section 4.

All the evaluated models were developed and tuned on the same training data set based on a stratified 5-fold cross-validation. Note that because of the low number of participants, a stratification by participants leads to highly biased folds and does not allow the classifiers to generalize well to unseen data. Therefore, we only employed a class-based stratified sampling of the instances (i.e., feature vectors) for the 5 folds. The results of the evaluation are summarized in Table 1.

As shown in Table 1, the results of the comparison of the different approaches show that the proposed data transformation for the generation of new feature vectors consistently improves a model's predictive performance across all methods, regardless of the type of classification technique.

As was expected, the baseline approach based on the GBDT method shows an excellent predictive performance. It is in a tie with the SVM classifier on the original feature vectors and clearly outperforms all other baseline classifiers on the transformed data set (i.e. the transformed feature vectors using the $\Phi_{\mathcal{F}}$-transformation). The GBTD classifier is only outperformed by the proposed ensemble model. Consisting of several GBDT models, the proposed approach consistently shows the best predictive performance. It outperforms all the baseline classifiers by a considerable margin, both in terms of the Accuracy and the ROC-AUC measure. The ROC curves for the compared models are depicted in Fig. 3. They were generated on the validation sets of the transformed training data and clearly show the robust performance of the proposed ensemble method over all classification thresholds. Fig. 4 presents confusion matrices for all approaches which are used in the present study. When compared to other approaches submitted in the context of the CogLoad@UbiComp 2020 Challenge, our method returned the best predictions on unseen data, thus winning first place in the competition. Also, in comparison to the previous work on the cognitive load estimation using sensor information from a wrist-band device (e.g. (Gjoreski et al., 2018; Schaule et al., 2018)), our approach showed superior results.

Table 1
The performance comparison between various ML models on the original (raw) and the transformed data set. We evaluate the models using two iterations of stratified 5-fold cross-validation (2x5cv) with the following performance metrics: accuracy (higher is better) and ROC-AUC (higher is better) and report the mean and ± std results. The top results for each data set are marked in bold.

	Original Data Set		Transformed Data Set	
Model	Accuracy	ROC-AUC	Accuracy	ROC-AUC
LR	0.52 ± 0.030	0.54 ± 0.037	0.63 ± 0.025	0.65 ± 0.015
kNN	0.52 ± 0.028	0.54 ± 0.047	0.59 ± 0.023	0.61 ± 0.043
SVM	0.58 ± 0.025	0.60 ± 0.018	0.62 ± 0.040	0.65 ± 0.038
RF	0.56 ± 0.034	0.60 ± 0.036	0.64 ± 0.036	0.67 ± 0.043
AdaBoost	0.57 ± 0.039	0.57 ± 0.037	0.61 ± 0.030	0.64 ± 0.049
MLP	0.57 ± 0.038	0.60 ± 0.037	0.58 ± 0.030	0.60 ± 0.049
GBDT	0.58 ± 0.027	0.60 ± 0.036	0.65 ± 0.031	0.68 ± 0.034
Ensemble Model	**0.59 ± 0.035**	**0.61 ± 0.036**	**0.66 ± 0.035**	**0.69 ± 0.035**

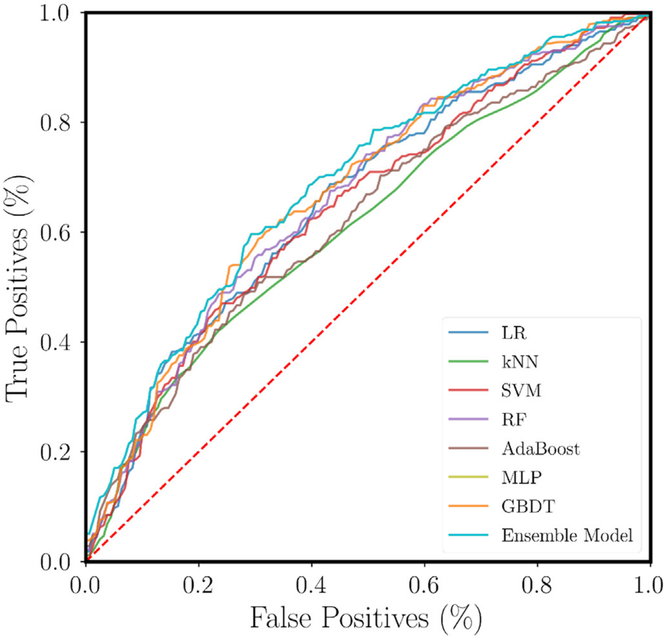


Fig. 3. ROC curves for all compared methods on the transformed data.

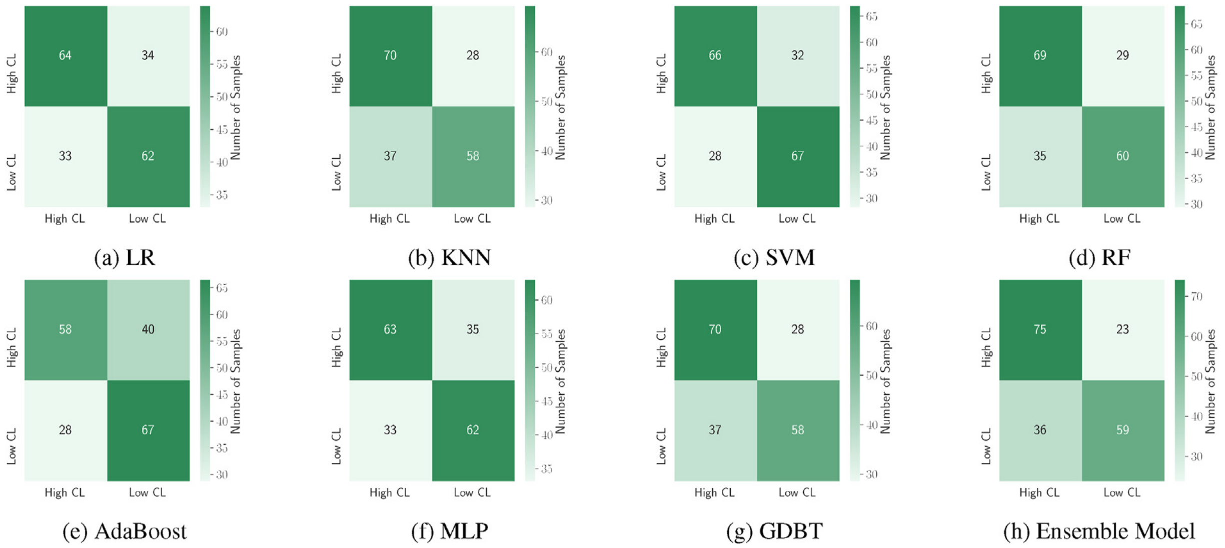
6.4. Variable importance

In Fig. 5, we present the feature importance analysis on the transformed data set. We employ the CancelOut neural layer proposed in (Borisov et al., 2019) and the Shapley additive explanations (SHAP) framework (Lundberg & Lee, 2017) to analyze the importance of the features for the prediction task at hand. Implemented as a neural layer, CancelOut was included as an additional layer to the MLP network. The SHAP analysis was conducted based on the GBDT model. The results depicted in Fig. 5 provide convincing evidence that the features produced by the standard deviation (std) function are among the most informative; there are three such features among the top-5 most important features. This is in line with our intuition and related work in the area of cognitive load detection which found that the variance of the measured physiological signals contains most of the information among the first three statistical moments of the (time-series) measures. The temperature sensor features, on the other hand, do not show a high relative importance. The RR- and HR-related features are the most informative according to the analysis, which is also quite intuitive, since the heart rate and its variability are known to be strongly correlated to cognitive load levels, e.g., (Gjoreski et al., 2018; Hughes et al., 2019; Wang & Guo, 2019).

Interestingly, the SHAP method assigns a higher importance to the skew-related features and lower importance to the mean-related features than the CancelOut method. Despite such differences, we can see that all the proposed transformations are valuable for the prediction of cognitive load from the available physiological signals.

7. Limitations and future work

Although our methods prove robust to subject-related bias and variability in the data, our results are based on a rather small data set derived from only 23 participants. Hence, further investigations on large data sets and higher variability are required. Of particular importance may be the investigation of inter-subject variability and whether or not and how the manifestation of cognitive load changes over time. Such an investigation requires, however, not only data collected in longitudinal studies, but also methods that are able to re-calibrate to subject-related



(a) LR (b) KNN (c) SVM (d) RF

(e) AdaBoost (f) MLP (g) GDBT (h) Ensemble Model

Fig. 4. Confusion matrices for baseline approaches and the proposed ensemble model on transformed test data set. As can be seen, our model shows superior performance among selected ML baseline in detection high cognitive load (High CL) by correctly identifying the state of the cognitive load; however, the SVM and Adaptive Boosting algorithms demonstrate the top performance in detecting the state where a subject has a regular or low cognitive load (Low CL). We use the same data splitting strategy as we did at the CogLoad@UbiComp 2020 competition, e. g in the test set, we have data for unseen subjects.

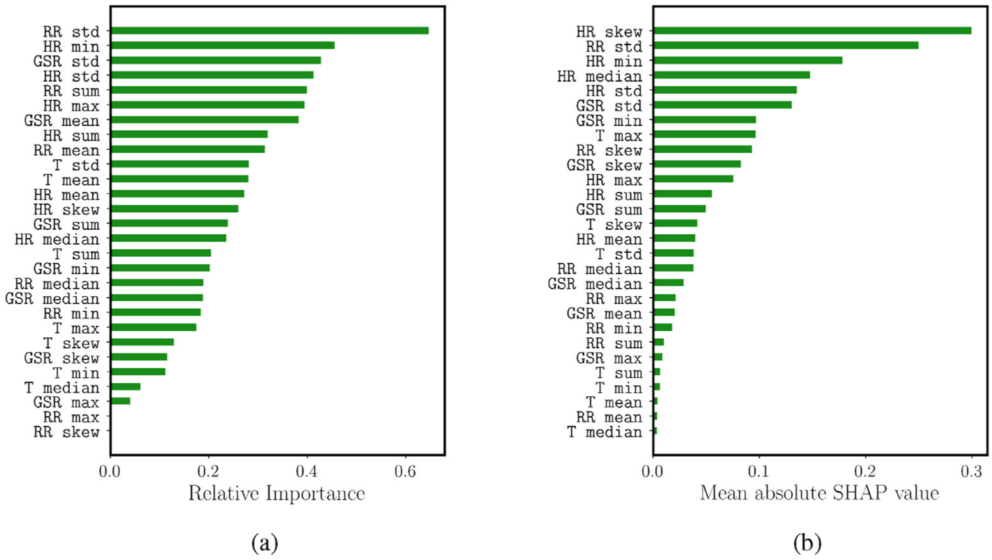


Fig. 5. Global variable importance analysis for the generated features using the CancelOut (a) and SHAP (b) algorithms.

characteristics.Regarding methodology, we also plan to investigate better ensemble strategies rather than simple mean aggregation. A better aggregation strategy for the ensemble step might improve the overall performance of our approach; one way is to rank models in the ensemble. A combination of the approaches might also help to estimate the cognitive load from a wearable device.

8. Conclusion

In this work, we presented an ensemble method for robust cognitive load detection based on physiological sensor data. With regard to its predictive power, our model outperformed the competing approaches, thus proving excellent robustness on unseen and small data. Furthermore, we showed that our proposed data transformation technique for the generation of new feature vectors from galvanic skin response, skin temperature, heart rate, and heart rate variability data notably improves the predictive power of machine learning techniques and might therefore be applicable to other areas of affective or cognition-aware computing. In contrast to the competing approaches, the effect of the input features on the model prediction is explainable, thus allowing a detailed analysis of cognitive load factors.

References

Abdelrahman, Y., Velloso, E., Dingler, T., Schmidt, A., & Vetere, F. (2017). Cognitive heat: Exploring the usage of thermal imaging to unobtrusively estimate cognitive load. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 1*, 1–20.

Amadieu, F., Van Gog, T., Paas, F., Tricot, A., & Mariné, C. (2009). Effects of prior knowledge and concept-map structure on disorientation, cognitive load, and learning. *Learning and Instruction, 19*, 376–386.

Appel, T., Scharinger, C., Gerjets, P., & Kasneci, E. (2018). Cross-subject workload classification using pupil-related measures. In *Proceedings of the 2018 ACM symposium on eye tracking research & applications* (pp. 1–8).

Appel, T., Sevcenko, N., Wortha, F., Tsarava, K., Moeller, K., Ninaus, M., Kasneci, E., & Gerjets, P. (2019). Predicting cognitive load in an emergency simulation based on behavioral and physiological measures. In *2019 international conference on multimodal interaction* (pp. 154–163).

Babu, M. D., JeevithaShree, D., Prabhakar, G., Saluja, K. P. S., Pashilkar, A., & Biswas, P. (2019). Estimating pilots' cognitive load from ocular parameters through simulation and in-flight studies. *Journal of Eye Movement Research, 12*.

Beatty, J., Lucero-Wagoner, B., et al. (2000). The pupillary system. *Handbook of psychophysiology, 2*.

Benedetto, S., Pedrotti, M., & Bridgeman, B. (2011). Microsaccades and exploratory saccades in a naturalistic environment. *Journal of Eye Movement Research, 4*, 1–10.

Bergstra, J., Yamins, D., & Cox, D. D. (2013). Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms. In *Proceedings of the 12th Python in science conference, Citeseer* (p. 20).

van Berkel, N., Exler, A., Gjoreski, M., Kolenik, T., Okoshi, T., Pejovic, V., Visuri, A., Voit, A., & Ubittention. (2020). *5th international workshop on smart & ambient notification and attention management*.

Bloomfield, P. (2004). *Fourier analysis of time series: An introduction*. John Wiley & Sons.

Borisov, V., Haug, J., & Kasneci, G. (2019). Cancelout: A layer for feature selection in deep neural networks. In *International conference on artificial neural networks* (pp. 72–83). Springer.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory* (pp. 144–152).

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32.

Bristow, D., Frith, C., & Rees, G. (2005). Two distinct neural effects of blinking on human visual processing. *NeuroImage, 27*, 136–145.

Chan, K. P., & Fu, A. W. C. (1999). Efficient time series matching by wavelets. In *Proceedings 15th international conference on data engineering (cat. No. 99CB36337)* (pp. 126–133). IEEE.

Chaovalit, P., Gangopadhyay, A., Karabatis, G., & Chen, Z. (2011). Discrete wavelet transform-based time series analysis and mining. *ACM Computing Surveys, 43*, 1–37.

Chen, S., & Epps, J. (2014). Using task-induced pupil diameter and blink rate to infer cognitive load. *Human-Computer Interaction, 29*, 390–413.

Chen, S., Epps, J., Ruiz, N., & Chen, F. (2011). Eye activity as a measure of human mental effort in hci. In *Proceedings of the 16th international conference on Intelligent user interfaces* (pp. 315–318).

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794).

De Silva, A.M., Leong, P.H., . Grammar-based feature generation for time-series prediction. Springer.

Debie, E., Rojas, R. F., Fidock, J., Barlow, M., Kasmarik, K., Anavatti, S., & Abbass, H. A. (2021). Multimodal fusion for objective assessment of cognitive workload: A review. *IEEE Transactions on Cybernetics, 51*(3), 1542–1555. https://doi.org/10.1109/TCYB.2019.2939399

Di Nocera, F., Camilli, M., & Terenzi, M. (2007). A random glance at the flight deck: Pilots' scanning strategies and the real-time assessment of mental workload. *Journal of Cognitive Engineering and Decision Making, 1*, 271–285.

Duchowski, A., Krejtz, K., Krejtz, I., Biele, C., Niedzielska, A., Kiefer, P., Martin, R., & Giannopoulos, I. (2018). *The index of pupillary activity: Measuring cognitive load vis-à-vis task difficulty with pupil oscillation*. https://doi.org/10.1145/3173574.3173856

Faure, V., Lobjois, R., & Benguigui, N. (2016). The effects of driving environment complexity and dual tasking on drivers' mental workload and eye blink behavior. *Transportation Research Part F: Traffic Psychology and Behaviour, 40*, 78–90.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences, 55*, 119–139.

Fridman, L., Reimer, B., Mehler, B., & Freeman, W. T. (2018). Cognitive load estimation in the wild. In *Proceedings of the 2018 CHI conference on human factors in computing systems* (pp. 1–9). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3173574.3174226.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis, 38*, 367–378.

Friedman, N., Fekete, T., Gal, Y. K., & Shriki, O. (2019). Eeg-based prediction of cognitive load in intelligence tests. *Frontiers in Human Neuroscience, 13*, 191.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1). New York: Springer series in statistics.

Fu, T.c. (2011). A review on time series data mining. *Engineering Applications of Artificial Intelligence, 24*, 164–181.

Fukuda, K., Stern, J. A., Brown, T. B., & Russo, M. B. (2005). Cognition, blinks, eye-movements, and pupillary movements during performance of a running memory task. *Aviation Space & Environmental Medicine, 76*, C75–C85.

Gao, X., Yan, H., & Sun, H.j. (2015). Modulation of microsaccade rate by task difficulty revealed through between-and within-trial comparisons. *Journal of Vision, 15*, 3–3.

Gardner, M. W., & Dorling, S. (1998). Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric Environment, 32*, 2627–2636.

Gjoreski, M., Kolenik, T., Knez, T., Luštrek, M., Gams, M., Gjoreski, H., & Pejović, V. (2020). Datasets for cognitive load inference using wearable sensors and psychological traits. *Applied Sciences, 10*, 3843.

Gjoreski, M., Luštrek, M., & Pejović, V. (2018). My watch says i'm busy: Inferring cognitive load with low-cost wearables. In *Proceedings of the 2018 ACM international joint conference and 2018 international symposium on pervasive and ubiquitous computing and wearable computers* (pp. 1234–1240).

Grinsted, A., Moore, J. C., & Jevrejeva, S. (2004). *Application of the cross wavelet transform and wavelet coherence to geophysical time series*.

Grubov, V., Badarin, A., & Maksimenko, V. (2020). Analysis of information perception and processing during long-term and intense cognitive load using combined eeg and nirs. In *2020 international conference nonlinearity, information and robotics (NIR)* (pp. 1–2). IEEE.

Hart, S. G., & Staveland, L. E. (1988). Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology* (Vol. 52, pp. 139–183). Elsevier.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer Science & Business Media.

He, J., & McCarley, J. S. (2010). Executive working memory load does not compromise perceptual processing during visual search: Evidence from additive factors analysis. *Attention, Perception, & Psychophysics, 72*, 308–316.

Hogervorst, M. A., Brouwer, A. M., & van Erp, J. B. F. (2014). Combining and comparing eeg, peripheral physiology and eye-related measures for the assessment of mental workload. *Frontiers in Neuroscience, 8*, 322. https://doi.org/10.3389/fnins.2014.00322, 25352774[pmid] http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4196537/.

Huang, W., Li, J., & Alem, L. (2018). Towards preventative healthcare: A review of wearable and mobile applications. *Data, Informatics and Technology: An Inspiration for Improved Healthcare*, 11–14.

Hughes, A. M., Hancock, G. M., Marlow, S. L., Stowers, K., & Salas, E. (2019). Cardiac measures of cognitive workload: A meta-analysis. *Human Factors, 61*, 393–414.

Inamdar, S., & Pomplun, M. (2003). Comparative search reveals the tradeoff between eye movements and working memory use in visual tasks. In *Proceedings of the annual meeting of the cognitive science society*.

Kagglecom. Kaggle competitions. URL https://www.kaggle.com/competitions.

Kahneman, D. (1973). *Attention and effort* (Vol. 1063). Citeseer.

Kasneci, E., Kübler, T., Broelemann, K., & Kasneci, G. (2017). Aggregating physiological and eye tracking signals to predict perception in the absence of ground truth. *Computers in Human Behavior, 68*, 450–455.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in neural information processing systems* (pp. 3146–3154).

Keshmiri, S., Sumioka, H., Yamazaki, R., & Ishiguro, H. (2017). A non-parametric approach to the overall estimate of cognitive load using nirs time series. *Frontiers in Human Neuroscience, 11*, 15.

Kosch, T., Hassib, M., Woźniak, P. W., Buschek, D., & Alt, F. (2018). Your eyes tell: Leveraging smooth pursuit for assessing cognitive workload. In *Proceedings of the 2018 CHI conference on human factors in computing systems* (pp. 1–13).

Kramer, A. E. (1990). *Physiological metrics of mental workload: A review of recent progress*.

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

Krejtz, K., Duchowski, A. T., Niedzielska, A., Biele, C., & Krejtz, I. (2018). Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze. *PloS One, 13*, Article e0203629.

Kübler, T. C., Kasneci, E., Rosenstiel, W., Schiefer, U., Nagel, K., & Papageorgiou, E. (2014). Stress-indicators and exploratory gaze for the analysis of hazard perception in patients with visual field loss. *Transportation Research Part F: Traffic Psychology and Behaviour, 24*, 231–243.

Kuncheva, L. I., & Whitaker, C. J. (2003). Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning, 51*, 181–207.

Kun, A. L., Palinko, O., Medenica, Z., & Heeman, P. A. (2013). *On the feasibility of using pupil diameter to estimate cognitive load changes for in-vehicle spoken dialogues*. INTERSPEECH.

Li, X., & De Cock, M. (2020). Cognitive load detection from wrist-band sensors. In *Adjunct proceedings of the 2020 ACM international joint conference on pervasive and ubiquitous computing and proceedings of the 2020 ACM international symposium on wearable computers* (pp. 456–461).

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 30, pp. 4765–4774). Curran Associates, Inc.. URL http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

Mäki-Marttunen, V., Hagen, T., & Espeseth, T. (2019). Task context load induces reactive cognitive control: An fmri study on cortical and brain stem activity. *Cognitive, Affective, & Behavioral Neuroscience, 19*, 945–965.

Marshall, S. P. (2000). Method and apparatus for eye tracking and monitoring Pipil dilation to evaluate cognitive activity. URL https://patentimages.storage.googleapis.com/pdfs/9171d27ab488a900c7db/US6090051.pdf.

Marshall, S. P. (2007). Identifying cognitive state from eye metrics. *Aviation Space & Environmental Medicine, 78*, B165–B175.

McDuff, D., Gontarek, S., & Picard, R. (2014). Remote measurement of cognitive stress via heart rate variability. In *2014 36th annual international conference of the IEEE engineering in medicine and biology society, IEEE* (pp. 2957–2960).

Mehler, B., Reimer, B., & Wang, Y. (2011). *A comparison of heart rate and heart rate variability indices in distinguishing single-task driving and driving under secondary cognitive workload*.

Mills, C., Fridman, I., Soussou, W., Waghray, D., Olney, A. M., & D'Mello, S. K. (2017). Put your thinking cap on: Detecting cognitive load using eeg during learning. In *Proceedings of the seventh international learning analytics & knowledge conference* (pp. 80–89).

Nourbakhsh, N., Chen, F., Wang, Y., & Calvo, R. A. (2017). Detecting users' cognitive load by galvanic skin response with affective interference. *ACM Transactions on Interactive Intelligent Systems (TiiS), 7*, 1–20.

Nourbakhsh, N., Wang, Y., Chen, F., & Calvo, R. A. (2012). Using galvanic skin response for cognitive load measurement in arithmetic and reading tasks. In *Proceedings of the 24th Australian computer-human interaction conference* (pp. 420–423).

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

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research, 12*, 2825–2830.

Pfleging, B., Fekety, D. K., Schmidt, A., & Kun, A. L. (2016). A model relating pupil diameter to mental workload and lighting conditions. In *Proceedings of the 2016 CHI conference on human factors in computing systems* (pp. 5776–5788).

Prabhakar, G., Mukhopadhyay, A., Murthy, L., Modiksha, M., Sachin, D., & Biswas, P. (2020). Cognitive load estimation using ocular parameters in automotive. *Transport Engineer, 2*, 100008.

Rafiqi, S., Wangwiwattana, C., Fernandez, E., Nair, S., & Larson, E. (2015). Work-in-progress, pupilware-m: Cognitive load estimation using unmodified smartphone cameras. In *2015 IEEE 12th international conference on mobile ad hoc and sensor systems* (pp. 645–650). https://doi.org/10.1109/MASS.2015.31

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*, 372.

Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review, 33*, 1–39.

Salfinger, A. (2020). Deep learning for cognitive load monitoring: A comparative evaluation. In *Adjunct proceedings of the 2020 ACM international joint conference on pervasive and ubiquitous computing and proceedings of the 2020 ACM international symposium on wearable computers* (pp. 462–467).

Schaule, F., Johanssen, J. O., Bruegge, B., & Loftness, V. (2018). Employing consumer wearables to detect office workers' cognitive load for interruption management. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2*, 1–20.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks, 61*, 85–117.

Schmiedek, F., Lövdén, M., & Lindenberger, U. (2014). A task is a task is a task: Putting complex span, n-back, and other working memory indicators in psychometric context. *Frontiers in Psychology, 5*, 1475.

Scholkopf, B., & Smola, A. J. (2018). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. Adaptive Computation and Machine Learning series.

Setz, C., Arnrich, B., Schumm, J., La Marca, R., Tröster, G., & Ehlert, U. (2010). Discriminating stress from cognitive load using a wearable eda device. *IEEE Transactions on Information Technology in Biomedicine, 14*, 410–417. https://doi.org/10.1109/TITB.2009.2036164

Sharma, K., Niforatos, E., Giannakos, M., & Kostakos, V. (2020). Assessing cognitive performance using physiological and facial features: Generalizing across contexts. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 4*, 1–41.

Shi, Y., Ruiz, N., Taib, R., Choi, E., & Chen, F. (2007). Galvanic skin response (gsr) as an index of cognitive load. In *CHI'07 extended abstracts on Human factors in computing systems* (pp. 2651–2656).

Siegenthaler, E., Costela, F. M., McCamy, M. B., Di Stasi, L. L., Otero-Millan, J., Sonderegger, A., Groner, R., Macknik, S., & Martinez-Conde, S. (2014). Task difficulty in mental arithmetic affects microsaccadic rates and magnitudes. *European Journal of Neuroscience, 39*, 287–294.

Stubbs, J. L., Corrow, S. L., Kiang, B., Panenka, W. J., & Barton, J. J. (2018). The effects of enhanced attention and working memory on smooth pursuit eye movement. *Experimental Brain Research, 236*, 485–495.

Sweller, J. (2011). Cognitive load theory. In *Psychology of learning and motivation* (Vol. 55, pp. 37–76). Elsevier.

international joint conference on pervasive, T.A., ubiquitous computing. (2020). Ubicomp2020. URL https://ubicomp.hosting.acm.org/ubicomp2020/.

Tervonen, J., Pettersson, K., & Mäntyjärvi, J. (2021). Ultra-short window length and feature importance analysis for cognitive load detection from wearable sensors. *Electronics, 10*, 613.

Van Gog, T., Paas, F., & Van Merriënboer, J. J. (2005). Uncovering expertise-related differences in troubleshooting performance: Combining eye movement and concurrent verbal protocol data. *Applied Cognitive Psychology, 19*, 205–221.

Van Orden, K. F., Limbert, W., Makeig, S., & Jung, T. P. (2001). Eye activity correlates of workload during a visuospatial memory task. *Human Factors, 43*, 111–121.

Verleysen, M., & François, D. (2005). The curse of dimensionality in data mining and time series prediction. In *International work-conference on artificial neural networks* (pp. 758–770). Springer.

Wang, C., & Guo, J. (2019). A data-driven framework for learners' cognitive load detection using ecg-ppg physiological feature fusion and xgboost classification. *Procedia computer science, 147*, 338–348.

Wang, Q., Yang, S., Liu, M., Cao, Z., & Ma, Q. (2014). An eye-tracking study of website complexity from cognitive load perspective. *Decision Support Systems, 62*, 1–10. https://doi.org/10.1016/j.dss.2014.02.007. URL http://www.sciencedirect.com/science/article/pii/S0167923614000402.

Wilson, G., Fullenkamp, P., & Davis, I. (1994). Evoked potential, cardiac, blink, and respiration measures of pilot workload in air-to-ground missions. *Aviation Space & Environmental Medicine, 65 2*, 100–105.