Computer Methods and Programs in Biomedicine 224 (2022) 106989

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

COLET: A dataset for COgnitive workLoad estimation based on eye-tracking

Emmanouil Ktistakis[a,b,1], Vasileios Skaramagkas[a,e,1,*], Dimitris Manousos[a], Nikolaos S. Tachos[c], Evanthia Tripoliti[d], Dimitrios I. Fotiadis[c], Manolis Tsiknakis[a,e]

[a] *Institute of Computer Science, Foundation for Research and Technology Hellas (FORTH), GR-700 13 Heraklion, Greece*
[b] *Laboratory of Optics and Vision, School of Medicine, University of Crete, GR-710 03 Heraklion, Greece*
[c] *Biomedical Research Institute, FORTH, GR-451 10, Ioannina, Greece and the Dept. of Materials Science and Engineering, Unit of Medical Technology and Intelligent Information Systems, University of Ioannina, GR-451 10, Ioannina, Greece*
[d] *Dept. of Materials Science and Engineering, Unit of Medical Technology and Intelligent Information Systems, University of Ioannina, GR-451 10, Ioannina, Greece*
[e] *Dept. of Electrical and Computer Engineering, Hellenic Mediterranean University, GR-710 04 Heraklion, Crete, Greece*

ARTICLE INFO

Article history:
Received 14 April 2022
Revised 2 June 2022
Accepted 28 June 2022

Keywords:
Cognitive workload
Workload classification
Eye movements
Machine learning
Eye-tracking
Affective computing

ABSTRACT

Background and Objective: The cognitive workload is an important component in performance psychology, ergonomics, and human factors. Publicly available datasets are scarce, making it difficult to establish new approaches and comparative studies. In this work, COLET-COgnitive workLoad estimation based on Eye-Tracking dataset is presented. Methods: Forty-seven (47) individuals' eye movements were monitored as they solved puzzles involving visual search activities of varying complexity and duration. The participants' cognitive workload level was evaluated with the subjective test of NASA-TLX and this score is used as an annotation of the activity. Extensive data analysis was performed in order to derive eye and gaze features from low-level eye recorded metrics, and a range of machine learning models were evaluated and tested regarding the estimation of the cognitive workload level. Results: The activities induced four different levels of cognitive workload. Multi tasking and time pressure have induced a higher level of cognitive workload than the one induced by single tasking and absence of time pressure. Multi tasking had a significant effect on 17 eye features while time pressure had a significant effect on 7 eye features. Both binary and multi-class identification attempts were performed by testing a variety of well-known classifiers, resulting in encouraging results towards cognitive workload levels estimation, with up to 88% correct predictions between low and high cognitive workload. Conclusions: Machine learning analysis demonstrated potential in discriminating cognitive workload levels using only eye-tracking characteristics. The proposed dataset includes a much higher sample size and a wider spectrum of eye and gaze metrics than other similar datasets, allowing for the examination of their relations with various cognitive states.

1. Introduction

The study of mental workload, also known as cognitive workload (CW), is a vital aspect in the areas of psychology, ergonomics, and human factors in order to understand and interpret the performance throughout an activity or process [1,2]. Despite the multitudinous and extended research in this area, each model defines and measures cognitive workload in a different way [3]. Although generally, when there is high task demand, there is also high mental workload, it is not always the case. Workload is not determined only by the tasks, each user has different characteristics and develops their own strategies in order to cope with the problem [4]. Different people with different experience and abilities can handle the same task or activity in a different way [5–7]. Therefore, cognitive workload can be described as the subjective experience of a task with certain characteristics [6].

Due to the diversity of criteria for CW, there are several methods for quantifying it. No sensor can provide an accurate picture of how an individual responds to an activity, thus CW is evaluated

[*] Corresponding author at: Institute of Computer Science, Foundation for Research and Technology Hellas (FORTH), GR-700 13 Heraklion, Greece.
E-mail address: mankti2@gmail.com (E. Ktistakis).
[1] Authors contributed equally

https://doi.org/10.1016/j.cmpb.2022.106989

indirectly, with the measurement of some variables, such as subjective ratings, performance features (e.g. accuracy, reaction time) and physiological data [3]. Various physiological features are being used, such as electrocardiac activity measures,ocular and brain measures [6]. Among them, the estimation of eye and gaze patterns and features seem to be able to efficiently determine workload levels [8,9]. Specifically, the number of fixations and fixation duration have been demonstrated to increase as cognitive load increases during mentally demanding activities such as surgical operations, simulated flight tasks and video games [10–12]. Saccades are the most often examined kind of eye movements in cognitive workload research [13]. The average peak saccadic velocity is seen to increase in a positive linear fashion as the cognitive effort increases [12,13]. Moreover, microsaccades are frequently employed to investigate cognitive processes [14]. Additionally, and given the much greater prevalence of short blinks under situations of high visual load, blink length and frequency are also sensitive indicators of cognitive effort [12,15], while blink rate alters significantly in proportion with a rise in the difficulty of a secondary task conducted in tandem during performing complex activities such as driving [16] or operating surgery [17]. Finally, the pupil area is significantly related to the user's current work difficulty [18,19], and eventually mean pupil diameter has been demonstrated to correlate positively with cognitive effort across a variety of activities [15,20,21].

Numerous studies have concentrated on determining cognitive effort purely on the basis of ocular characteristics for various activities, highlighting the need of further research [22]. The majority of them provide binary categorization findings, indicating a high or low level of cognitive workload, with some obtaining highly accurate results [21,23–25]. However, there are just a few published attempts that focus on multi-class classification (high/medium/low), and the resulting performance is inferior [23,26]. The three-class classification is gaining popularity during the last years [27]. Finally in a recent study [28], participants performed activities of mental attentional capacity with six levels of difficulty and predictive models were developed based on metrics associated with activity difficulty, reaction time and eye movements. The outcomes demonstrated that machine learning models may accurately predict performance, with response times and level of difficulty serving as the most accurate predictors.

Although eye movements have proven to be useful indicators of cognitive processes [9], only few authors have focused on the development of relevant databases. Amongst those available, MAMEM datasets (Phase 1 and Phase 2) [29] blend multimodal biosignals and eye tracking data collected within the context of human-computer interaction. The datasets contain eye tracking data from 34 people (18 able-bodied and 16 with motor impairments), as well as electroencephalography (EEG), galvanic skin response (GSR), and heart rate (HR) signals. The data were collected during engagement with a specially built interface for online browsing and manipulating multimedia material, as well as during fictitious mobility activities.

The EGTEA Gaze+ dataset [30] comprises almost 28 hours of video footage from 86 separate sessions including 32 people completing seven distinct food preparation activities. There are movies, eye-tracking data, action annotations, and hand masks included in the dataset. It is a supplement to the previously released GTEA Gaze+ dataset [31]. Moving beyond the dataset, the authors propose a novel deep model for joint gaze estimation and action recognition in first person vision.

In USC CRCNS Dataset [32], the authors used abrupt transitions to convert continuous video clips into clip parts (jump cuts). 16 participants had their saccadic motions recorded as objective behavioral markers of attentional choices. By assessing the agreement between human attentional selection and prediction produced by a neurally grounded computational model, they were able to measure the usage of perceptual memory across viewing circumstances and across time. Additionally, MIT CVCL Search Model Database [33] comprises of eye-tracking recordings from 14 participants while performing person detection activities. The ground-truth eye movement data were used to evaluate three computational models for search guidance based on saliency, target features, and scene context respectively.

In this work, we present COLET: A Dataset for COgnitive workLoad estimation based on Eye-Tracking. The dataset explores the possibility to analyze CW levels induced by visual search puzzles along with secondary tasks performed from different users. The collection contains eye and gaze movement recordings of 47 participants and their performance scores when solving CAPTCHA-like puzzles related to visual search activities. Visual search is a very important feature of human activity and many studies have evaluated cognitive workload during visual search [34,35]. The CAPTCHA puzzle was chosen because it is an activity that is usually met in real world. The recorded signals contain a number of metrics related to gaze positions, blinks and pupil characteristics, enabling for the extraction and analysis of a broad variety of eye features, such as fixations and saccades. After the conclusion of each activity, ratings from individuals in relation to a simplified version of the NASA activity load index (NASA TLX) tool [36] were collected. NASA TLX is a multidimensional valid measure of CW that is commonly used bye experts [4,37]. Moreover, this work offers an extensive description of the eye and gaze data processing techniques applied from the initiation of the recording process, including the computational pipeline for the extraction of fixation, blink, saccade and pupil related features. Finally, a variety of well-known machine learning classifiers is implemented in order to examine the potential of eye and gaze information provided in COLET for CW estimation, demonstrating important findings regarding the models' prediction capability.

To our knowledge, this is one of very few public databases with eye-tracking data obtained from mentally demanding visual search puzzles, that consists of such a high number of participants, while additionally providing with a detailed data processing methodology towards the utilization of a wide spectrum of machine learning methods for CW estimation. COLET is publicly available [38] and can contribute towards the development and evaluation of modern human-computer interaction and recommendation systems.

2. Methods

In this Section, we describe the protocol followed for generating the dataset and every material that was used in the study.

2.1. Participants

Exclusion criteria for the participants included: any known ocular disease, spectacle-corrected binocular visual acuity in 80 cm worse than 0.10 logMAR (0.8 decimal acuity equivalent), clinically significant abnormal phorias (see Section 2.2 for more details).

Fifty six (56) individuals volunteered for the study and nine of them were excluded: Seven due to the exclusion criteria and two because of poor quality recordings. Thus, analysis was performed for the remaining forty-seven (47) participants (26 female, 21 male). Their mean age was 32±8 years (range: 18–47 years), their mean education level was 17±2 years (range:12-21 years) and their mean binocular visual acuity at 80 cm was -0.10 ± 0.08 logMAR (range: 0.10-(-0.29) logMAR).

The experimental protocol was submitted (104/12-1-2021/03-2-2021) and approved (110/12-02-2021) by the Ethical Committee of the Foundation for Research and Technology Hellas (FORTH).

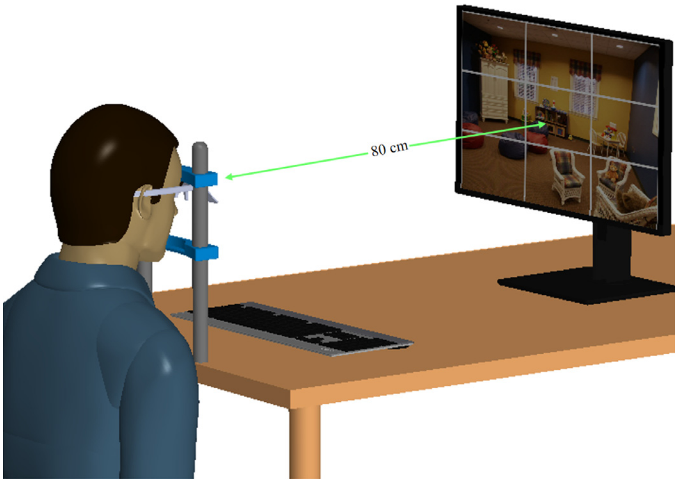


Fig. 1. Graphical representation of the experimental setup.

2.2. Materials and setup

A set of 21 images of indoor scenes was chosen from the free database Indoor Scene Recognition [39]. A grid was added to each image, thus dividing it into nine (9) equal squares as it is shown in Fig. 3. Each image was selected so as a specific object was present on some of the 9 squares, thus looking like a CAPTCHA puzzle. The participants were asked to spot and indicate the squares of the puzzle with a certain object. This consisted the Main Task.

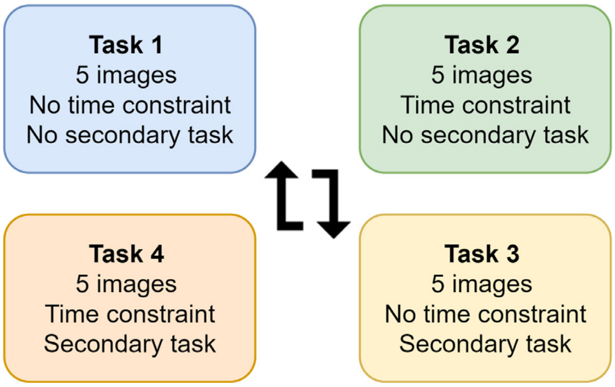
An interference task was also posed to the participants as a secondary task, during which participants were asked to count aloud and backwards from 1000 by subtracting 4.

The images were presented on a computer screen (LCD, 24″, 1280x720) at 80cm distance from the participant as it is shown in Fig. 1.

Eye-tracking measurements were recorded using the Pupil Labs ”Pupil Core” eye-tracker. Recordings were binocular with 240 Hz sampling frequency, accuracy 0.60° and precision 0.02. All measurements were performed with the participants seated on a chair with their head stabilized by means of a chin and head rest to minimize head movements. Furthermore, an API in Python programming environment was developed to communicate with the eye-tracker and to enable/disable the eye and gaze data acquisition procedure according to the stage of the experiment as well as to identify possible errors of the participant during the procedure as described in the next paragraphs. The API was also responsible for collecting the eye and gaze data after the end of each trial and saving them with the proper labelling. The eye and gaze raw data were saved in.csv format. Finally, a Graphical User Interface (GUI) in Python programming environment was designed, in order to give the proper instructions to the participants regarding the experimental procedure and the sequence of the activities. Using this GUI, the participants could navigate through the experimental phases and also cancel the procedure in case they decided to.

Integrity of participants' vision was evaluated in terms of visual acuity and binocular coordination. Visual acuity was measured with the European-wide standardized logMAR charts [40]. Binocular coordination (i.e. any clinically significant phorias) was evaluated with the cover test.

Recordings were performed under controlled, photopic lighting conditions, which were achieved with the room lights on. Illuminance at cornea when screen was off, was 400 lx and when on, in blank screen, it was 450 lx.



Task 1
5 images
No time constraint
No secondary task

Task 2
5 images
Time constraint
No secondary task

Task 3
5 images
No time constraint
Secondary task

Task 4
5 images
Time constraint
Secondary task

Fig. 2. Two-by-two factorial design of the experimental study.

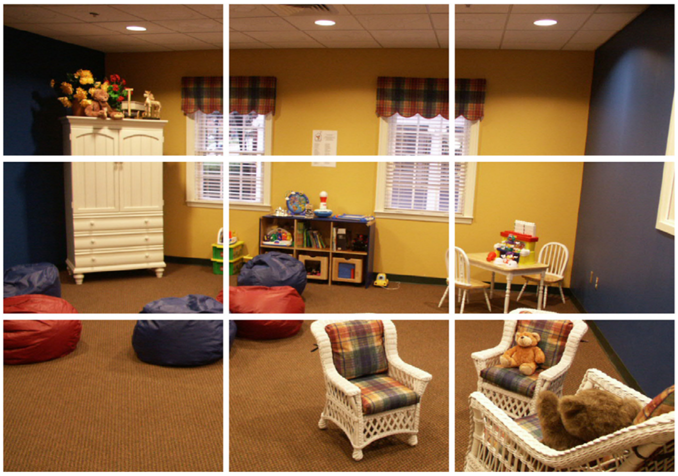


Fig. 3. A sample trial/image of the CAPTCHA test. Instructions: 'Choose the squares in which pouffes are located”.

2.3. Experimental procedure

In the beginning, all participants read and signed an Information Consent Form. Subsequently, a binocular visual acuity test at distance of 80 cm and a stereopsis test were conducted. Afterwards, they were led to the experiment room. All the necessary measures for the protection of the participants and the research team from the SARS-CoV-2 pandemic and expansion of the coronavirus were applied.

Following that, participants were instructed to complete some demographics (age, education level) on a computer screen. A test with a random image was conducted next in order the participant to familiarize with the process. After the test, the main part of the study commenced.

The study used a two-by-two factorial design, with the two factors being Time Constraint (with or without) and Tasking (single or multi tasking). Time Constraint was imposed by instructing participants to finish the assignment ”as quickly as possible,” whereas ”no time constraint” was introduced by instructing participants to complete the activity ”at a comfortable pace”. Single tasking consisted of the aforementioned Main Task and multi tasking consisted of the Secondary along with the Main Task. The interaction of the two factors resulted in the establishment of four experimental activities; Activity1: no time pressure and single task, Activity2: time pressure and single task, Activity3: no time pressure and multi tasking, Activity4: time pressure and multi tasking (Fig. 2).

Each activity consisted of 5 images/trials. For each participant, the activities were presented in a randomised order and for each activity 5 images were selected randomly from the pool of images, in order to avoid any learning and fatigue effects. Only after the presentation of all 5 images of an activity, the images of the next activity were presented. Each activity was presented only once to each participant. Between the activities, there was a two minute break.

At the end of each activity the participants were asked to complete a simplified version of the NASA-TLX questionnaire, a subjective workload assessment tool [36].

An image from the ones used in the experimental procedure is shown in Fig. 3. Throughout the activity completing procedure, a member of the research group monitored the gaze tracker's output on a second screen in case any anomalies in the recordings occurred or the participant needed further assistance. At any time during the study, participants could request that the process be stopped and their data deleted.

2.4. Cognitive workload assessment

A variety of measures to evaluate cognitive workload have been used and they can be divided in four categories; subjective measures, performance measures, psychophysiological measures and analytical measures [3].

In this study, subjective, performance and physiological measures were used to evaluate cognitive workload.

CW assessment was evaluated only for the Main task, as the Secondary task was used as a means to induce CW to the participants while performing the Main task.

2.4.1. Subjective measures

NASA-TLX consists of six subscales and originally it derives an overall workload score based on a weighted average of ratings on these subscales. A simplified version of NASA-TLX, called NASA RTLX was utilised, proposed by Byers et al. [41] and has been used widely ever since [42,43]. In NASA RTLX, the pairwise comparisons of the subscales that are used in the original NASA-TLX for weighing, are omitted. The subscales are the following: Mental Demand, Physical Demand, Temporal Demand, Performance, Effort and Frustration.

An example of the NASA-TLX questions is the one regarding mental demand: "How much mental and perceptual activity was required? Was the activity easy or demanding, simple or complex?" The subscales are rated within a 100-point range with 5-point steps. Apart from the ratings of the six subscales, the mean value of the six ratings was also evaluated. The higher the mean value, the higher the experienced cognitive workload [36].

2.4.2. Performance measures

During all trials and activities, the number of mistakes and missed correct squares was measured. Additionally, the time that was needed to complete a trial was measured and it is referred to as Reaction Time (RT). An attempt to combine speed and error is the Inverse Efficiency Score (IES) [44] which is widely used. For a given participant, IES is computed by mean RT in a particular condition divided by the percentage of correct answers (PC) [45]. PC was calculated as the number of the correct answers divided by the sum of the correct, the wrong and the missed answers.

2.4.3. Physiological measures

Physiological measures were derived from the eye tracker raw data. They are in total 28 measures and they are fixation, saccade, blink and pupil related, including skewness, kurtosis and coefficient of variation (CV) for every eye feature. Skewness is a measure of symmetry of a distribution, while kurtosis is a measure of whether a distribution is heavy-tailed or light-tailed relative to a normal distribution. Coefficient of variation shows the extent of variability in relation to the mean of the population and is a dimensional number. It is defined as the ratio of the standard deviation to the mean of a population. For ease of reading, it will be referred to as Variation. A detailed presentation of the features studied, is given in Section 3.

2.5. Data analysis methodology

The computational procedure that was followed to calculate and evaluate the eye and gaze related features from the raw data acquired by the gaze tracking device is described in detail below.

2.6. Raw data processing

When collecting data, some noise is typically present due to eye blinking and failure to capture corneal reflections (i.e., signal loss) [46]. The gaze tracker's output includes the gaze positions (x,y coordinates), blink timings (start and end times), and pupil diameter in mm. These measures include a variety of sources of noise, including the eye-tracker and the participants. Filtering and denoising are used to eliminate this undesired volatility in eye movement data [47].

The raw gaze coordinates in normalized pixels form are converted to degrees of visual angle, and the instantaneous sample-to-sample gaze movement between two consecutive gaze locations is determined, resulting in the computation of the angular velocity at the specified sampling frequency F_s. To decrease velocity noise, we used a five-tap FIR velocity filter with its form modified in response to a defined velocity peak value during a saccade [48]. Owing to its longer sampling window, the filter is more effective at signal smoothing (anti-aliasing).

2.7. Fixation and saccade detection

In this work, fixations and saccades are identified based on the Velocity-Threshold Identification (I-VT) algorithm [49] due to its superiority when considering sample-by-sample comparisons [50]. Additionally, we introduced an additional minimum time criterion to assess the duration of the fixations. According to the method, a defined velocity threshold determines a gaze point as a fixation or saccade. Then, consecutive fixation points are collapsed into fixation groups based on the duration threshold. In the I-VT algorithm the velocity threshold for saccade detection was set to 45 deg./sec, as in [50]. In addition, the minimum fixation duration threshold were determined at 55 msec [51].

2.8. Pupil and blink detection

The effect of a certain factor on pupil size is hard to evaluate, since pupil diameter and its variation is highly dependent on multiple factors that need to remain fixed, such as lighting conditions [52–54] and the adapting field size [55,56]. In our study, the luminance of each image may be a factor of pupil size change.

The current study's experimental design was chosen to minimize this impact as much as possible. Initially, the lighting settings of the room were configured to be photopic, ensuring that the effect of brightness shifts of the images was minimal. For the same purpose, the screen was positioned at a distance of 80 cm away from the participant.

In order to evaluate whether eventually the influence of the changes of the image luminance was sufficiently low, linear regression analysis between mean pupil diameter of each participant and the V component of the HSV color space of each image was carried out. The HSV color space stands for Hue, Saturation and Value

Table 1
Chosen estimators and the tested hyperparameters.

Estimator	Hyperparameters tested
GNB	loss: [deviance, exponential], criterion: [friedman_mse, squared_error, mse, mae], min_samples_split: [1, 2, 5, 10]
RF	max_depth: [None, 5, 10, 20, 30], max_features: [auto, sqrt], min_samples_leaf: [1, 2, 4], min_samples_split: [2, 4, 6], n_estimators: [10, 100, 200, 500, 1000, 1200]
SVM	decision_function_shape: [ovo, ovr], gamma: [scale, auto]
EGB	criterion: [friedman_mse, squared_error, mse, absolute_error], loss: [deviance, exponential], max_depth: [None, 5, 10, 20, 30], max_features: [auto, sqrt, log2], min_samples_leaf: [1, 2, 4], min_samples_split: [2, 4, 6], n_estimators: [10, 100, 200, 500, 1000, 1200]
k-NN	algorithm: [auto, ball_tree, kd_tree, brute], leaf_size: [10, 20, 30, 40, 50], n_neighbors: [1, 3, 5, 10, 20, 30], p: [1, 2, 3, 4, 5], weights: [uniform, distance]
NB	alpha: [1.0, 2.0, 5.0, 10.0]
LR	penalty: [l1, l2, elasticnet, none], solver: [newton-cg, lbfgs, liblinear, sag, saga], multi_class: [auto, ovr, multinomial]
DT	criterion: [gini, entropy], max_depth: [None, 5, 10, 20, 30], max_features: [auto, sqrt, log2], min_samples_leaf: [1, 2, 4], min_samples_split: [2, 4, 6], splitter: [best, random]

and is an alternative to RGB color model. The V component describes the brightness of a color. The idea was based on a recent attempt to remove the movie luminance effect by extracting the estimated pupil diameter based on the V component of the HSV color space, from the recorded pupil diameter [57]. Among the 47 participants, only 2 showed correlation; one moderate ($r = 0.442$, $p = 0.034$) and one strong ($r = 0.635$, $p = 0.003$). The results were considered satisfactory since pupil diameter did not seem to correlate strongly with the V component. Thus, all analysis on pupil diameter was carried out with the pupil diameter obtained from the eye tracker.

2.9. Feature selection and model training

Variables that are measured at different scales do not contribute equally to the model fitting and model learned function and this can often caused biased behavior [58]. Thus, in order to deal with this potential problem, we performed feature-wise normalization prior to model fitting stage. Specifically, after the features were extracted, we transformed them by scaling each feature to a range between 0 and 1 using the MinMaxScaler function. Furthermore, we constructed a correlation matrix showing correlation coefficients between the variables as a preliminary data interpretation tool and an input into a more advanced analysis, in our case the analysis of variance (ANOVA). Based on the ANOVA repeated measures analysis performed in Section 3.1, we derived the most dominant features for every classification attempt described in the next paragraphs.

In total, 8 classifiers were trained and tested during the classification procedure. More specifically: Gaussian Naive Bayes (GNB), Random Forest (RF), Linear Support Vector Machine (SVM), Ensemble Gradient Boosting (EGB), K-Nearest Neighbor (k-NN), Bernoulli Naive Bayes (NB), Logistic Regression (LR) and Decision Trees (DT) [59]. We selected a range of well-known and widely used classifiers to study their behavior in comparison to the respective literature. To fine tune the hyperparameters of each classifier we performed a RandomSearch iterating 1000 times through training data to find the combination of parameters that maximizes the overall performance and accuracy. The selected classifiers along with the tested hyperparameters values are shown in Table 1. Finally, we split the data into training and testing, with the number of the test data being 20% of the total number of examples, according to the Pareto principle [60]. We evaluated the models using the metrics of accuracy and f1-score. Furthermore, we validated the models using a k-fold cross-validation ($k = 5$). The models were built using Python (3.8 Python packages, version numpy 1.17.2, sklearn 0.21.3, pandas 0.25.1.

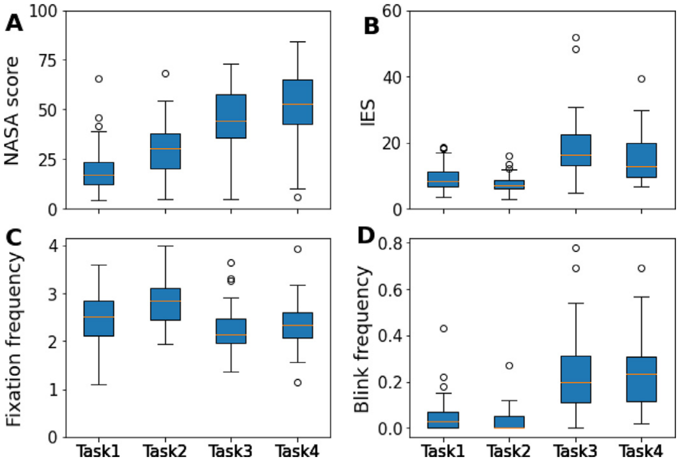


Fig. 4. Box plots for A) Mean NASA score B) IES C) fixation frequency and D) blink frequency in each activity.

Specifically and for the machine learning analysis, we attempt to identify relations between fixation, saccade, blink and pupil related eye features and the CW activities and levels. Therefore, the four activities that the participants engaged with during the experimental procedure are noted as **A1, A2, A3, A4** for activities 1, 2, 3 and 4, respectively as shown in Fig. 2. Moreover, an additional machine learning analysis was performed based on the subjective annotation as extracted from the mean NASA RTLX scores per activity given by the participants. In the same manner, the outcome measure regarding CW levels can have three values: low, medium and high cognitive workload based on the average workload that indicates the intensity of the perceived workload [61]. Therefore, we define high, medium and low CW levels for workload score between 50–100, 30–49 and 0–29 respectively [62]. Additionally, the divided CW instances low, medium and high are marked as class **C0, C1** and **C2**, respectively.

3. Results

In this section, the results of the study are presented. The statistical analysis is initially shown and then the Machine Learning analysis which has been performed to identify any relation between eye features and cognitive workload activities and levels.

3.1. Statistical analysis

3.1.1. Cognitive workload induction

Table 2 shows mean values (±SD) of all NASA subscales in all activities. A one-way repeated measures ANOVA determined that both all NASA RTLX subscale scores and mean NASA score showed statistically significant difference among the activities. Post hoc analysis with a Bonferroni adjustment revealed that mean NASA was statistically significantly different between all pairs of activities ($p < 0.014$) and it was getting gradually higher when moving from activity 1 to activity 4 (Fig. 4A). Mental demand was statistically significantly higher in activity 2 compared to activity 1 (11.0 (95% CI, 4.0 to 18.0), $p < 0.0001$) and even higher in activity 3 (30.2 (95% CI, 19.9 to 40.5), $p < 0.0001$). Mental demand in activity 4 was not statistically significantly different from the mental demand in activity 3 (6.0 (95% CI, −1.1 to 13.2), $p = 0.143$). Temporal demand was higher in activity 2 compared to activity 1 (22.5 (95% CI, 10.5 to 34.6), $p < 0.0001$) and higher in activity 4 com-

Table 2
NASA RTLX scores (±SD).

Subscale	Act. 1	Act. 2	Act. 3	Act. 4
Mental Demand	25.6 (±18.2)	36.7 (±23.8)	66.9 (±23.4)	72.9 (±19.8)
Physical Demand	16.0 (±18.2)	19.6 (±15.5)	26.5 (±27.0)	27.3 (±25.9)
Temporal Demand	19.0 (±14.8)	41.5 (±27.6)	37.3 (±22.8)	59.9 (±28.0)
Performance	16.6 (±20.0)	23.2 (±17.4)	43.0 (±23.5)	45.6 (±24.3)
Effort	26.7 (±20.0)	36.7 (±23.0)	63.8 (±21.8)	66.8 (±21.8)
Frustration	12.5 (±15.6)	17.4 (±17.7)	31.6 (±22.6)	40.4 (±27.9)
Mean	19.4 (±11.9)	29.2 (±13.6)	44.8 (±15.1)	52.2 (±17.1)

Act.1: no time pressure-single tasking, Act.2: time pressure-single tasking, Act.3: no time pressure-multi tasking, Act.4: time pressure-multi tasking.

Table 3
Results of two-way repeated measures ANOVA in subjective and performance measures.

	Tasking		Time		Interaction
Measure	p	Dif. (multi-single)	p	Dif. (with-without)	p
Mental demand	**0.000**	38.8	**0.000**	8.5	0.209
Physical demand	**0.006**	9.1	0.363		0.614
Temporal demand	**0.000**	18.4	**0.000**	22.6	0.981
Performance	**0.000**	24.4	0.073		0.439
Effort	**0.000**	33.6	**0.003**	6.5	0.083
Frustration	**0.000**	21.1	**0.007**	6.9	0.369
Mean NASA	**0.000**	24.2	**0.000**	8.6	0.445
Mistakes	**0.008**	0.14	**0.047**	0.11	0.056
Reaction time	**0.000**	29.0	**0.000**	−11.0	0.776
IES	**0.000**	8.4	**0.004**	−2.5	0.294

Two-way repeated measures ANOVA with Bonferroni adjustment for subjective and performance measures. p-value and mean difference of all features between two levels of two factors: tasking (multi or single) and time pressure (with or without). Statistically significant differences ($p < 0.05$) are in bold.

pared to activity 3 (22.6 (95% CI, 12.1 to 33.1), $p < 0.0001$), as expected. Performance and Frustration were statistically significantly different only between single (1 and 2) and double (3 and 4) activities, while Effort was statistically significantly different between all pairs apart from activity 3 and activity 4 (3.0 (95% CI, −5.5 to 11.5), $p = 1.000$). Finally, post hoc analysis did not show any statistically significant difference in Physical Demand between any pair of activities ($p > 0.102$).

Repeated measures ANOVA also showed that the mean number of mistakes per activity ($F(3, 141) = 4.354$, $P = 0.006$), the total time needed to complete a activity (Reaction Time, RT) ($F(2.371, 111.445) = 49.878$, $P < 0.0001$) and the Inverse Efficiency Score (IES) ($F(2.062, 96.905) = 40.443$, $P < 0.0001$) differed statistically significantly among the different activities. Post hoc analysis with a Bonferroni adjustment revealed that the number of mistakes was statistically significantly lower in activity 1 compared to the the number of mistakes in activities 3 and 4 ($p < 0.026$), while no difference was found between the rest of the activities. RT was statistically significantly lower in activity 2 compared to activity 1 (10.355 (95% CI, 4.46 to 16.25) sec, $p < 0.0001$) and lower in activity 4 compared to activity 3, without reaching significance though (11.74 (95% CI, −0.78 to 24.25) sec, $p = 0.078$). RT was also statistically significantly higher in activities 3 and 4 compared to activities 1 and 2 ($p < 0.0001$). Post hoc analysis showed that the Inverse Efficiency Score (IES) was statistically significantly different among all pairs, apart from between activity 3 and activity 4 (3.285 (95% CI, −0.804 to 7.374) sec, $p = 0.191$) (Fig. 4B).

Based on all subjective and performance measures, apart from Temporal Demand, it becomes evident that cognitive workload is increased as one moves from activity 1 to activity 4. Among all measures, mean NASA score seems to be the measure that can better distinguish among the activities. Thus, from now on, cognitive workload (CW) will be considered as follows:

$$CW_{activity1} < CW_{activity2} < CW_{activity3} < CW_{activity4}.$$

Further analysis was conducted with two-way repeated measures ANOVA with the two factors being Tasking (single or multi) and Time (with or without time constraint). The results are shown in Table 3. There was a significant main effect of Tasking on all subjective and performance measures that were investigated and a significant main effect of Time on 8 out of 10 features. No significant interaction between Tasking and Time was observed.

3.1.2. Eye feature analysis

A two-way repeated measures ANOVA was performed for all eye features. The two factors were asking (single or multi) and Time (with or without time constraint). There was a significant main effect of asking on seventeen (17) and a significant main effect of Time on seven (7) out of 28 features. There was also a significant interaction between asking and Time in 3 features. Mean values of all features across activities are presented in Table 4. Post hoc analysis with a Bonferroni adjustment was also performed and mean differences of the features between two levels of each factor are presented in Table 5. Indicatively, box plots of fixation frequency and blink frequency are shown in Fig. 4C and 4 D.

3.2. Machine learning analysis

The results of the classification study using COLET are presented in Tables 6 and 7. Each of the tables contains the classes of the respective classification attempt, the sample size for each class, and the three best performing classifiers in terms of their accuracy. The classification results are evaluated in terms of accuracy, precision, recall and f1-score. Moreover, the best classification algorithm for each classification attempt is highlighted in bold.

From the results presented in Table 6, activity 2 (A2) which required from the participants to complete the activity ”as soon as

Table 4
Mean values and standard deviations of all features for the four activities.

	Activity 1		Activity2		Activity 3		Activity 4	
Feature	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.
Fixation frequency (fix./sec)	2.50	0.53	2.80	0.48	2.24	0.47	2.35	0.50
Fixation duration (msec)	272.65	65.97	244.43	46.06	269.32	63.63	254.13	42.21
Variation	0.85	0.14	0.79	0.15	0.87	0.14	0.86	0.13
Skewness	2.49	0.87	2.26	0.67	2.49	0.84	2.57	1.04
Kurtosis	8.65	6.94	6.49	4.45	8.78	8.32	9.74	11.13
Saccade frequency (sac./sec)	1.71	0.85	1.85	0.90	3.40	2.56	3.35	2.55
Saccade amplitude (deg.)	14.10	0.92	13.96	0.99	14.15	1.45	14.06	1.29
Variation	0.12	0.06	0.10	0.05	0.23	0.11	0.22	0.10
Skewness	−0.39	1.55	−0.26	1.52	0.39	1.70	−0.07	1.70
Kurtosis	4.62	6.81	3.13	6.88	4.88	6.10	5.42	7.12
Saccade velocity (deg./sec)	146.41	68.55	132.91	68.85	261.13	103.09	267.90	113.34
Variation	0.77	0.30	0.61	0.28	0.75	0.19	0.75	0.21
Skewness	2.06	1.14	1.95	1.07	0.88	1.12	0.86	1.30
Kurtosis	5.08	6.06	4.84	5.28	0.72	3.68	1.10	4.67
Peak saccade velocity (deg./sec)	216.89	89.26	204.14	90.82	350.49	114.24	357.25	126.08
Variation	0.77	0.21	0.66	0.20	0.70	0.18	0.69	0.21
Skewness	1.64	1.10	1.60	1.16	0.51	1.08	0.45	1.16
Kurtosis	3.27	5.41	3.81	5.82	−0.14	3.02	0.04	2.61
Saccade duration (msec)	15.33	3.27	15.39	2.61	19.05	4.59	19.16	3.51
Variation	0.97	0.37	0.92	0.42	1.25	0.35	1.38	0.70
Skewness	3.00	1.39	2.88	1.93	2.98	1.04	3.39	2.23
Kurtosis	12.41	10.51	12.96	16.56	11.82	9.46	17.93	39.80
Blink frequency (blinks/sec)	0.05	0.07	0.03	0.05	0.24	0.17	0.24	0.17
Blink duration (msec)	205.54	48.35	200.31	47.53	229.00	40.48	212.29	31.97
Pupil diameter (mm)	3.49	0.60	3.66	0.61	3.80	0.69	3.82	0.71
Variation	0.05	0.04	0.05	0.03	0.07	0.04	0.06	0.03
Skewness	−0.30	0.53	−0.41	0.59	−0.16	0.82	−0.46	0.86
Kurtosis	1.02	4.25	1.04	2.22	2.34	3.30	3.70	6.16

Fixation duration, saccade amplitude and saccade duration are median values. Saccade velocity, peak saccade velocity, blink duration and pupil diameter are mean values. Fix./sec: number of fixations per second, sac./sec: number of saccades per second, deg.: degrees of visual angle during a saccadic movement, deg./sec: degrees of visual angle during a saccadic movement per second, blinks/sec: number of blinks per second.

Table 5
Results of two-way repeated measures ANOVA.

	asking		Time		Interaction
Feature	p	Dif. (multi-single)	p	Dif. (with-without)	p
Fixation frequency	**0.000**	−0.355	**0.000**	0.222	**0.026**
Fixation duration	0.638		**0.000**	−23.365	0.175
Fixation duration Variation	**0.003**	0.045	**0.022**	−0.039	0.208
Saccade frequency	**0.000**	1.624	0.752		0.579
Saccade amplitude Variation	**0.000**	0.118	0.146		0.392
Saccade velocity	**0.000**	125.456	0.648		0.214
Saccade velocity Variation	0.244		**0.024**	−0.072	**0.019**
Saccade velocity Skewness	**0.000**	−1.155	0.561		0.879
Saccade velocity Kurtosis	**0.000**	−4.093	0.885		0.721
Peak saccade velocity	**0.000**	144.184	0.714		0.288
Peak saccade velocity Variation	0.458		**0.016**	−0.059	**0.035**
Peak saccade velocity Skewness	**0.000**	−1.167	0.646		0.792
Peak saccade velocity Kurtosis	**0.000**	−3.648	0.503		0.692
Saccade duration	**0.000**	3.758	0.841		0.938
Saccade duration Variation	**0.000**	0.373	0.448		0.143
Blink frequency	**0.000**	0.201	0.473		0.067
Blink duration	**0.011**	21.901	0.652		0.232
Pupil diameter	**0.000**	0.254	**0.012**	0.106	0.149
Pupil diameter Variation	**0.001**	0.015	0.443		0.930
Pupil diameter Skewness	0.616		**0.042**	−0.196	0.190
Pupil diameter Kurtosis	**0.004**	1.994	0.246		0.100

Two-way repeated measures ANOVA with Bonferroni adjustment. p-value and mean difference of all features between two levels of two factors: tasking (multi or single) and time pressure (with or without). Statistically significant differences ($p < 0.05$) are in bold.

possible" but without counting backwards was distinguished from A3 and A4 at percentages of 93 and 98%, respectively, demonstrating the effect of the secondary task (backwards counting) to the differentiation of the activities. In the same manner, in A1/A3 and A1/A4 binary classification problems, the accuracy rate of GNB and k-NN classifiers was found to be 81 and 86%, respectively. Additionally, time pressure factor played a critical role in the discrimination among the first (A1) and second (A2) experimental activities with NB achieving 80% accuracy.

Interestingly, the backwards counting which was common for experimental activities A3 and A4 seems to have outweighed their difference which was the time pressure, as the highest accuracy percentage was observed from DT classifier at 60%. Moreover, the addition of the secondary task lead to the effective identification

Table 6
Superior algorithms in classifying the four experimental activities for the induction of cognitive workload.

Classes	Sample size	Superior classifiers	Acc.	Pr.	Rec.	f1
A1/A2	47/47	ENS	.75	.75	.74	.75
		RF	.79	.80	.80	.80
		NB	.80	.79	.78	.78
A1/A3	47/47	ENS	.78	.79	.79	.79
		SVM	.75	.74	.75	.74
		GNB	.81	.79	.79	.79
A1/A4	47/47	**k-NN**	.86	.85	.86	.85
		LR	.85	.85	.84	.85
		DT	.80	.80	.79	.80
A2/A3	47/47	**SVM**	.93	.92	.91	.92
		GNB	.91	.91	.90	.91
		RF	.90	.91	.91	.91
A2/A4	47/47	**SVM**	.98	.98	.98	.97
		GNB	.96	.94	.95	.95
		RF	.98	.97	.98	.97
A3/A4	47/47	**DT**	.60	.60	.60	.60
		RF	.55	.53	.54	.54
		k-NN	.55	.55	.56	.54
A1,2/A3,4	94/94	ENS	.88	.87	.88	.87
		RF	.88	.88	.87	.88
		k-NN	.90	.89	.89	90
A1/A2/	47/47/	**NB**	.60	.59	.59	.60
A3/A4	47/47	LR	.55	.54	.54	.54
		GNB	.52	.52	.54	.51

Table 7
Superior algorithms in classifying three levels of cognitive workload.

Classes	Sample size	Superior classifiers	Acc.	Pr.	Rec.	f1
C1/C2	71/67	**ENS**	.72	.72	.75	.72
		SVM	.69	.67	.68	.69
		GNB	.71	.71	.71	.72
C1/C3	71/50	SVM	.84	.82	.83	.82
		GNB	.88	.86	.86	.86
		RF	.84	.84	.85	.84
C2/C3	67/50	ENS	.58	.58	.59	.57
		RF	.54	.54	.54	.55
		DT	.62	.61	.61	.60
C1,2/C3	138/50	ENS	.72	.70	.71	.71
		RF	.72	.71	.69	.70
		DT	.74	.73	.74	.73
C1/C2/C3	71/67/50	RF	.54	.52	.51	.52
		GNB	.59	.59	.58	.59
		LR	.51	.50	.50	.50

of A3 and A4 instances from A1 and A2 (no secondary task) with 90% prediction accuracy achieved from k-NN classifier, whereas RF and ENS accuracy rates remaining close enough. On the contrary, the multi-class problem regarding the synchronous classification of all four activities with the NB classifier, decreased the accuracy to 60%.

The results of our attempt to predict the levels of CW based on the mean scores of the NASA-TLX are presented in Table 7. Almost 9 out of 10 examples of low (C1) and high (C3) CW were classified correctly by the GNB classifier. On the contrary, the identification of medium levels of CW (C2) from the rest two classes (high and low) proved particularly challenging for the classifiers. Specifically, C1 and C2 examples were classified correctly from the ENS model at 72% accuracy rate, while the 62% of C2 and C3 cases were predicted properly by the DT classifier.

The DT classifier was proven superior in correctly identifying high CW (C3) from the other two levels based on the NASA-TLX mean score with 74% accuracy.

The last classification problem is related to the classification of three levels of CW; high, medium and low. The GNB was proved to be the most efficient in terms of accuracy reaching up to 59% correct predictions. Overall, GNB and DT models seemed to be able to identify correctly the three levels of CW, high, medium and low, however the insertion of the medium class decreased significantly the accuracy percentage.

4. Discussion

Despite the major contribution of eye-tracking databases for CW quantification to the scientific community as described in the Section 1, each of the databases has distinct drawbacks. The methodologies used suffer from a limited number of available eye and gaze measurements. This is especially critical when examining relations between eye movements and cognitive states, since some measures, such as blink duration and saccadic velocity, play a vital role in the estimation of increased cognitive workload [9]. Furthermore, the above mentioned datasets do not primarily target to study the alterations of ocular movements in relation to cognitive load variations. To address these issues, we presented an eye-tracking dataset to be used for the analysis of cognitive workload levels and comprises of eye and gaze recordings signals from 47 participants, where each participant engaged in visual search related activities. Each activity differs from the others in terms of the existence of time constraint and a secondary task and is rated from the participants based on the NASA RTLX workload index.

In this work, we presented a dataset comprising of eye movement features gathered as each of the participants solved visual search puzzles and conducted supplementary activities, later translated in terms of cognitive workload levels. Despite the considerable contribution of analogous databases to the research community, our proposed dataset includes a much higher sample size and a wider spectrum of eye and gaze metrics, allowing for the examination of their relations with various cognitive states. In parallel, the dataset is annotated using not only the individuals' NASA RTLX scores, but also the activities in which they participated.

The statistical analysis revealed that the activities induced different levels of cognitive workload. Both subjective and performance measures reveal that multi tasking and time pressure have induced a higher level of CW than the one induced by single tasking and absence of time pressure. Two-way repeated measures ANOVA showed that multi tasking had a significant effect on 17 eye features while time pressure had a significant effect on 7 eye features. Statistically significant interaction between tasking and time pressure was observed in three features; fixation frequency, saccade velocity variation and peak saccade velocity variation. Further analysis with paired sample t-tests showed that although time pressure affected fixation frequency, the effect was much stronger when there was also multi tasking. Paired sample t-tests also showed that saccade velocity variation ered peak saccade velocity variation were affected significantly by time pressure only when there was single tasking and that multi tasking did have a significant effect on saccade velocity variation when there was time pressure. The low effect of time pressure compared to that of multi tasking can be, at least partially, attributed to the way time pressure was induced to the participants. This topic is discussed further on, in the limitations of the study. Fixation frequency and pupil diameter seem to be the most sensitive features as they exhibit a significant effect of both multi tasking and time pressure. Fixation frequency decreases in multi tasking and increases with time pressure, while pupil diameter increases both with multi tasking and time pressure (Table 5).

Additionally, we evaluated a range of eye parameters including fixations, saccades, blinks, and pupil size, as well as the capabilities of numerous machine learning models in a variety of categorization scenarios. Our findings corroborate earlier research and reveal that cognitive workload has an influence on eye movements and pupillary responses. Specifically and in line with the ideas of

[21,23–25], cognitive workload levels may be recognized successfully using only eye-tracking characteristics. Furthermore, our findings extend beyond prior studies such as [26], revealing considerable advances in terms of accuracy and highlighting the importance of continuing research in this field. Additionally, we established a substantial correlation between ocular characteristics and the four experimental activities, demonstrating the possibility of developing a cognitive workload detection system with a high degree of discretization capability.

Regarding the effort of quantifying the CW activities, the highest success rate was observed during the binary classification between A2 and A4, achieving 98% accuracy, while A2 is effectively distinguished also from A4. However, the classification attempts between activities which both included or not included the secondary task, resulted in considerable loss in model performance, particularly when distinguishing between A3 and A4. The strong effect of backwards counting in model accuracy is confirmed with the effective identification of A1 and A2 from A3 and A4. Additionally, it was challenging for the models to identify separately the four activities at a satisfactory level. At this point, the importance of the secondary task during an activity must be highlighted. Specifically, the models that classified between activities without secondary task and time pressure factor (i.e. A2) and activities with a secondary task and time pressure factor(i.e. A4) or between activities without secondary task and no time pressure factor (i.e. A1) and activities with a secondary task and no time pressure factor(i.e. A3) provided with better results (accuracy, precision, recall and f1-score) than most of the models that attempted to classify between categories which both included -or not- the secondary task independently of the existence of time pressure factor such as A3 and A4. This finding is in line with the NASA - RTLX scores where the secondary tasks show higher increases than time constrains.

In terms of estimating cognitive workload levels, both binary and multi-class identification tests produced encouraging results, with up to 88% correct predictions between low and high CW with the GNB classifier. Furthermore, the C2 class had a substantial effect on the models' performance resulting in accuracy decrease. Finally, results indicated that the GNB model emboldens the further investigation of classification between three or more CW levels by the addition of extra number of samples.

Despite the aforementioned advantages of our work that make the dataset an important contribution to the scientific community, the empirical results reported herein should be considered in the light of some limitations. First of all, an analysis of CW levels based not on different activities but on trials themselves, would result in a much greater sample size, thus providing the opportunity to exploit different machine learning methods for CW identification purposes. This was not applied in the current study because of some certain aspects of the study design. First of all, the duration of each trial was not long enough to have robust gaze pattern alterations among trials.The mean values of our eye features would be highly affected by any extreme values, as there would be just a few samples in each trial. Secondly, the CW annotation that we used, was based on the NASA-TLX questionnaire, that was filled by the participants after the end of each activity, thus we do not have NASA-TLX score for each trial. Furthermore, it seems that the CW induced by the time pressure factor was not as high as expected. We assume that the simple instructions posed to the participants to finish the assignment "as quickly as possible", were not capable to induce high CW. The use of a visible countdown timer instead of a simple instruction, would most probably be more appropriate. Another factor that may have affected our results is the size of the stimuli. Smaller images or greater number of squares may had induced higher cognitive workload, as the size of the stimuli affects visual search [63].

Our results encourage the development of a cognitive workload identification system with high discretization ability. The results presented in this article based on the data provided by COLET, demonstrate the potential of utilizing eye and gaze markers for discriminating between the various CW levels, while reinforcing the imperative need for future research. Towards this direction, the dataset is made available to the public and we firmly encourage other researchers and academics to test their methods and algorithmic approaches on this highly challenging database. New models as well as the applicability of deep learning methods have to be investigated in order to create a model for precisely estimating CW levels with high efficiency. This model could later be used in practical applications, especially in working environments leading to the development of a decision-making system that assesses the mental strain of the worker and provides with recommendations and advice for better time and workload management.

Declaration of Competing Interest

Conflict of Interest and Authorship Conformation Form Please check the following as appropriate:

- All authors have participated in (a) conception and design, or analysis and interpretation of the data; (b) drafting the article or revising it critically for important intellectual content; and (c) approval of the final version.
- This manuscript has not been submitted to, nor is under review at, another journal or other publishing venue.
- The authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript
- The following authors have affiliations with organizations with direct or indirect financial interest in the subject matter discussed in the manuscript:

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 826429 (Project: SeeFar). This paper reflects only the author's view and the Commission is not responsible for any use that may be made of the information it contains.

References

[1] F.N. Biondi, A. Cacanindin, C. Douglas, J. Cort, Overloaded and at work: investigating the effect of cognitive workload on assembly task performance, Hum. Factors 63 (5) (2021) 813–820, doi:10.1177/0018720820929928.

[2] I.P. Bodala, N.I. Abbasi, Y. Sun, A. Bezerianos, H. Al-Nashash, N.V. Thakor, Measuring vigilance decrement using computer vision assisted eye tracking in dynamic naturalistic environments, in: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2017, pp. 2478–2481, doi:10.1109/EMBC.2017.8037359. ISSN: 1558-4615

[3] B. Xie, G. Salvendy, Review and reappraisal of modelling and predicting mental workload in single- and multi-task environments, Work Stress 14 (2000) 74–99, doi:10.1080/026783700417249.

[4] F.P. da Silva, Mental workload, task demand and driving performance: what relation? Procedia - Social Behav. Sci. 162 (2014) 310–319, doi:10.1016/j.sbspro.2014.12.212. XVIII Congreso Panamericano de Ingenierḑa de Trnsito, Transporte y Logḑstica (PANAM 2014)

[5] T. Csipo, A. Lipecz, P. Mukli, D. Bahadli, O. Abdulhussein, C.D. Owens, S. Tarantini, R.A. Hand, V. Yabluchanska, J.M. Kellawan, F. Sorond, J.A. James, A. Csiszar, Z.I. Ungvari, A. Yabluchanskiy, Increased cognitive workload evokes greater neurovascular coupling responses in healthy young adults, PLoS ONE 16 (5) (2021) e0250043, doi:10.1371/journal.pone.0250043.

[6] R.L. Charles, J. Nixon, Measuring mental workload using physiological measures: a systematic review, Appl. Ergon. 74 (2019) 221–232, doi:10.1016/j.apergo.2018.08.028.

[7] E. Debie, R. Fernandez Rojas, J. Fidock, M. Barlow, K. Kasmarik, S. Anavatti, M. Garratt, H.A. Abbass, Multimodal fusion for objective assessment of cognitive workload: a review, IEEE Trans. Cybern. 51 (3) (2021) 1542–1555, doi:10.1109/TCYB.2019.2939399.

[8] D. Tao, H. Tan, H. Wang, X. Zhang, X. Qu, T. Zhang, A systematic review of physiological measures of mental workload, Int. J. Environ. Res. Public Health 16 (2019), doi:10.3390/ijerph16152716.

[9] V. Skaramagkas, G. Giannakakis, E. Ktistakis, D. Manousos, I. Karatzanis, N. Tachos, E.E. Tripoliti, K. Marias, D.I. Fotiadis, M. Tsiknakis, Review of eye tracking metrics involved in emotional and cognitive processes, IEEE Rev. Biomed. Eng. (2021), doi:10.1109/RBME.2021.3066072. 1–1

[10] G.G. Menekse Dalveren, N.E. Cagiltay, Insights from surgeons eye-movement data in a virtual simulation surgical training environment: effect of experience level and hand conditions, Behav. Inf. Technol. 37 (5) (2018) 517–537, doi:10.1080/0144929X.2018.1460399.

[11] X. He, L. Wang, X. Gao, Y. Chen, The eye activity measurement of mental workload based on basic flight task, in: IEEE International Conference on Industrial Informatics (INDIN), 2012, pp. 502–507, doi:10.1109/INDIN.2012.6301203.

[12] R. Mallick, D. Slayback, J. Touryan, A.J. Ries, B.J. Lance, The use of eye metrics to index cognitive workload in video games, in: Proceedings of the 2nd Workshop on Eye Tracking and Visualization, ETVIS 2016, Institute of Electrical and Electronics Engineers Inc., 2017, pp. 60–64, doi:10.1109/ETVIS.2016.7851168.

[13] I.P. Bodala, Y. Ke, H. Mir, N.V. Thakor, H. Al-Nashash, Cognitive workload estimation due to vague visual stimuli using saccadic eye movements, in: 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2014, Institute of Electrical and Electronics Engineers Inc., 2014, pp. 2993–2996, doi:10.1109/EMBC.2014.6944252.

[14] I.P. Bodala, S. Kukreja, J. Li, N.V. Thakor, H. Al-Nashash, Eye tracking and EEG synchronization to analyze microsaccades during a workload task, in: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2015, pp. 7994–7997, doi:10.1109/EMBC.2015.7320247. ISSN: 1558-4615

[15] M.I. Ahmad, I. Keller, D.A. Robb, K.S. Lohan, A framework to estimate cognitive load using physiological data, Pers. Ubiquitous Comput. (2020) 1–15, doi:10.1007/s00779-020-01455-7.

[16] T. Čegovnik, K. Stojmenova, G. Jakus, J. Sodnik, An analysis of the suitability of a low-cost eye tracker for assessing the cognitive load of drivers, Appl. Ergon. 68 (2018) 1–11, doi:10.1016/j.apergo.2017.10.011.

[17] R. Bednarik, J. Koskinen, H. Vrzakova, P. Bartczak, A.P. Elomaa, Blink-based estimation of suturing task workload and expertise in microsurgery, in: Proceedings - IEEE Symposium on Computer-Based Medical Systems, vol. 2018, Institute of Electrical and Electronics Engineers Inc., 2018, pp. 233–238, doi:10.1109/CBMS.2018.00048.

[18] M. Pomplun, S. Sunkara, Pupil dilation as an indicator of cognitive workload in human-computer interaction, 2003.

[19] E.H. Hess, J.M. Polt, Pupil size as related to interest value of visual stimuli, Science (1960) 349–350, doi:10.1126/science.132.3423.349.

[20] M. Nakayama, Y. Hayakawa, Relationships between oculo-motor mesures as task-evoked mental workloads during an manipulation task, in: Proceedings of the International Conference on Information Visualisation, vol. 2019, Institute of Electrical and Electronics Engineers Inc., 2019, pp. 170–174, doi:10.1109/IV.2019.00037.

[21] G. Prabhakar, A. Mukhopadhyay, L. Murthy, M. Modiksha, D. Sachin, P. Biswas, Cognitive load estimation using ocular parameters in automotive, Transp. Eng. 2 (2020) 100008, doi:10.1016/j.treng.2020.100008.

[22] B.A. Wilbanks, E. Aroke, K.M. Dudding, Using eye tracking for measuring cognitive workload during clinical simulations: literature review and synthesis, Comput. Inf. Nurs. 39 (9) (2021) 499–507, doi:10.1097/CIN.0000000000000704.

[23] V. Skaramagkas, E. Ktistakis, D. Manousos, N.S. Tachos, E. Kazantzaki, E.E. Tripoliti, D.I. Fotiadis, M. Tsiknakis, Cognitive workload level estimation based on eye tracking: a machine learning approach, in: 2021 IEEE 21st International Conference on Bioinformatics and Bioengineering (BIBE), 2021, pp. 1–5, doi:10.1109/BIBE52308.2021.9635166.

[24] X. Liu, T. Chen, G. Xie, G. Liu, Contact-free cognitive load recognition based on eye movement, J. Electr. Comput. Eng. 2016 (2016) 1–8, doi:10.1155/2016/1601879.

[25] C. Wu, J. Cha, J. Sulek, T. Zhou, C. Sundaram, J. Wachs, D. Yu, Eye-tracking metrics predict perceived workload in robotic surgical skills training, Hum. Factors 62 (2019) 001872081987454, doi:10.1177/0018720819874544.

[26] J. Chen, Q. Zhang, L. Cheng, X. Gao, L. Ding, A cognitive load assessment method considering individual differences in eye movement data, in: IEEE International Conference on Control and Automation, ICCA, vol. 2019, IEEE Computer Society, 2019, pp. 295–300, doi:10.1109/ICCA.2019.8899595.

[27] M. Plechawska, M. Tokovarov, M. Kaczorowska, D. Zapala, A three-class classification of cognitive workload based on eeg spectral data, Appl. Sci. 9 (2019) 5340, doi:10.3390/APP9245340.

[28] V. Bachurina, S. Sushchinskaya, M. Sharaev, E. Burnaev, M. Arsalidou, A machine learning investigation of factors that contribute to predicting cognitive performance: difficulty level, reaction time and eye-movements, Decis. Support Syst. 155 (2022) 113713, doi:10.1016/j.dss.2021.113713.

[29] S. Nikolopoulos, K. Georgiadis, F. Kalaganis, G. Liaros, I. Lazarou, K. Adam, A. Papazoglou-Chalikias, E. Chatzilari, V. Oikonomou, P. Petrantonakis, I. Kompatsiaris, C. Kumar, R. Menges, S. Staab, D. Müller, K. Sengupta, S. Bostantjopoulou, Z. Katsarou, G. Zeilig, M. Plotnik, A. Gotlieb, S. Fountoukidou, J. Ham, D. Athanasiou, A. Mariakaki, D. Comandicci, E. Sabatini, W. Nistico, M. Plank, A multimodal dataset for authoring and editing multimedia content: the MAMEM project, Data Brief (2017), doi:10.5281/zenodo.834154.

[30] Y. Li, M. Liu, J.M. Rehg, In the eye of beholder: joint learning of gaze and actions in first person video, in: Lecture Notes in Computer Science (LNCS) (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 11209, Springer Verlag, 2018, pp. 639–655, doi:10.1007/978-3-030-01228-1_38.

[31] A. Fathi, Y. Li, J.M. Rehg, Learning to recognize daily actions using gaze, in: Lecture Notes in Computer Science, vol. 7572, Springer, Berlin, Heidelberg, 2012, pp. 314–327, doi:10.1007/978-3-642-33718-5_23.

[32] R. Carmi, L. Itti, The role of memory in guiding attention during natural vision, J. Vis. 6 (9) (2006), doi:10.1167/6.9.4. 4–4

[33] K.A. Ehinger, B. Hidalgo-Sotelo, A. Torralba, A. Oliva, Modelling search for people in 900 scenes: a combined source model of eye guidance, Vis. Cogn. 17 (6–7) (2009) 945–978, doi:10.1080/13506280902834720. PMID: 20011676

[34] L.R. Enders, R.J. Smith, S.M. Gordon, A.J. Ries, J. Touryan, Gaze behavior during navigation and visual search of an open-world virtual environment, Front. Psychol. 12 (2021), doi:10.3389/fpsyg.2021.681042.

[35] M. Ranchet, J. Morgan, A.E. Akinwuntan, H. Devos, Exploring the cognitive workload during a visual search task in Parkinson's Disease, 2019, (MDS 2019, International Congress of Parkinson's Disease and Movement Disorders), Poster - MDS 2019, International Congress of Parkinson's Disease and Movement Disorders, Nice, FRANCE, 22-/09/2019 - 26/09/2019, https://hal.archives-ouvertes.fr/hal-02384125.

[36] S.G. Hart, L.E. Staveland, Development of NASA-TLX (task load index): results of empirical and theoretical research, Adv. Psychol. 52 (1988) 139–183, doi:10.1016/S0166-4115(08)62386-9.

[37] M. Mohammadian, H. Parsaei, H. Mokarami, R. Kazemi, Cognitive demands and mental workload: a filed study of the mining control room operators, Heliyon 8 (2) (2022) e08860, doi:10.1016/j.heliyon.2022.e08860.

[38] E. Ktistakis, V. Skaramagkas, D. Manousos, N.S. Tachos, E. Tripoliti, D.I. Fotiadis, M. Tsiknakis, COLET: A Dataset for Cognitive workLoad estimation based on Eye-Tracking, 2022, Type: dataset. 10.5281/ZENODO.5913227

[39] A. Quattoni, A. Torralba, Recognizing Indoor Scenes, 2010, pp. 413–420, doi:10.1109/CVPR.2009.5206537.

[40] S. Plainis, Y. Orphanos, M.K. Tsilimbaris, A modified ETDRS visual acuity chart for european-wide use, Optom. Vis. Sci. 84 (2007) 647–653.

[41] J.C. Byers, A. Bittner, S. Hill, Traditional and raw task load index (TLX) correlations: are paired comparisons necessary?, Taylor & Francis, 1989, pp. 481–485.

[42] L.D.J. Shiber, D.N. Ginn, A. Jan, J.T. Gaskins, S.M. Biscette, R. Pasic, Comparison of industry-leading energy devices for use in gynecologic laparoscopy: articulating enseal versus ligasure energy devices, J. Minim. Invasive Gynecol. 25 (2018) 467–473.e1, doi:10.1016/J.JMIG.2017.10.006.

[43] M. Georgsson, Nasa RTLX as a novel assessment tool for determining cognitive load and user acceptance of expert and user-based usability evaluation methods, Eur. J. Biomed. Inf. 16 (2020), doi:10.24105/EJBI.2020.16.2.14.

[44] J. Townsend, G. Ashby, Methods of modeling capacity in simple processing systems, Cognit. Theory 3 (1978).

[45] R. Bruyer, M. Brysbaert, Combining speed and accuracy in cognitive psychology: is the inverse efficiency score (IES) a better dependent variable than the mean reaction time (rt) and the percentage of errors (pe)? Psychol. Belg. 51 (2011) 5–13, doi:10.5334/PB-51-1-5.

[46] S. Taran, V. Bajaj, Emotion recognition from single-channel EEG signals using a two-stage correlation and instantaneous frequency-based filtering method, Comput. Methods Programs Biomed. 173 (2019) 157–165, doi:10.1016/j.cmpb.2019.03.015.

[47] S. Mejia-Romero, J. Eduardo Lugo, D. Bernardin, J. Faubert, An effective filtering process for the noise suppression in eye movement signals, in: K. Ray, K.C. Roy, S.K. Toshniwal, H. Sharma, A. Bandyopadhyay (Eds.), Proceedings of International Conference on Data Science and Applications, Springer, Singapore, 2021, pp. 33–46, doi:10.1007/978-981-15-7561-7-2.

[48] A.T. Duchowski, Eye movement analysis, in: Eye Tracking Methodology: Theory and Practice, Springer London, London, 2003, pp. 111–128, doi:10.1007/978-1-4471-3750-4-9.

[49] D.D. Salvucci, J.H. Goldberg, Identifying fixations and saccades in eye-tracking protocols, in: ETRA '00: Proceedings of the 2000 Symposium on Eye Tracking Research & Applications, Association for Computing Machinery, New York, NY, USA, 2000, pp. 71–78, doi:10.1145/355017.355028.

[50] R. Andersson, L. Larsson, K. Holmqvist, M. Stridh, M. Nystrm, One algorithm to rule them all? An evaluation and discussion of ten eye movement event-detection algorithms, Behav. Res. Methods 49 (2) (2016) 616–637, doi:10.3758/S13428-016-0738-9.

[51] S.M.K.A. Zaidawi, M.H.U. Prinzler, J. Lührs, S. Maneth, An extensive study of user identification via eye movements across multiple datasets, CoRR (2021) arXiv:2111.0590.

[52] C.J. Ellis, The pupillary light reflex in normal subjects, Br. J. Ophthalmol. 65 (1981) 754–759, doi:10.1136/BJO.65.11.754. https://pubmed.ncbi.nlm.nih.gov/7326222/

[53] C. Aracena, S. Basterrech, V. Snášel, J.D. Velásquez, Neural networks for emotion recognition based on eye tracking data, in: 2015 IEEE International Conference on Systems, Man, and Cybernetics, 2015, pp. 2632–2637.

[54] G. Pignoni, S. Komandur, F. Volden, Accounting for effects of variation in luminance in pupillometry for field measurements of cognitive workload, IEEE Sens. J. 21 (5) (2021) 6393–6400, doi:10.1109/JSEN.2020.3038291.

[55] D.A. Atchison, C.C. Girgenti, G.M. Campbell, J.P. Dodds, T.M. Byrnes, A.J. Zele, Influence of field size on pupil diameter under photopic and mesopic light levels, Clin. Exp. Optometry 94 (2011) 545–548, doi:10.1111/J.1444-0938.2011.00636.X. https://pubmed.ncbi.nlm.nih.gov/21929524/

[56] A.B. Watson, J.I. Yellott, A unified formula for light-adapted pupil size, J. Vis. 12 (2012) 12–12, doi:10.1167/12.10.12.

[57] P. Tarnowski, M. Kołodziej, A. Majkowski, R.J. Rak, Eye-tracking analysis for emotion recognition, Comput. Intell. Neurosci. 2020 (2020) 1687–5265, doi:10.1155/2020/2909267.

[58] P. Agrawal, H.F. Abutarboush, T. Ganesh, A.W. Mohamed, Metaheuristic algorithms on feature selection: asurvey of one decade of research (2009–2019), IEEE Access 9 (2021) 26766–26791, doi:10.1109/ACCESS.2021.3056407.

[59] A. Rcz, D. Bajusz, K. HĘberger, Multi-level comparison of machine learning classifiers and their performance metrics, Molecules 24 (15) (2019), doi:10.3390/molecules24152811.

[60] A. Gholamy, V. Kreinovich, O. Kosheleva, Why 70/30 or 80/20 relation between training and testing sets: a pedagogical explanation, 2018.

[61] I.B. Turksen, N. Moray, K. Fuller, A linguistic rule-based expert system for mental workload, in: H.-J. Bullinger, H.J. Warnecke (Eds.), Toward the Factory of the Future, Springer Berlin Heidelberg, Berlin, Heidelberg, 1985, pp. 865–875.

[62] I. Nur, H. Iskandar, R. Ade, The measurement of nurses' mental workload using NASA-TLX method (a case study), Malays. J. Public Health Med. (2020).

[63] F. Krause, H. Bekkering, J. Pratt, O. Lindemann, Interaction between numbers and size during visual search, Psychol. Res. 81 (2017) 664–677, doi:10.1007/s00426-016-0771-4.