Cognitive workload level estimation based on eye tracking: A machine learning approach

Vasileios Skaramagkas*, Emmanouil Ktistakis*†, Dimitris Manousos*, Nikolaos S. Tachos‡,
Eleni Kazantzaki*, Evanthia E. Tripoliti§, Dimitrios I. Fotiadis‡§ and Manolis Tsiknakis*¶
*Institute of Computer Science, Foundation for Research and Technology Hellas (FORTH),
GR-700 13 Heraklion, Crete, Greece, Email: vskaramag@ics.forth.gr
†Laboratory of Optics and Vision, School of Medicine, University of Crete,
GR-710 13 Heraklion, Crete, Greece
‡Dept. of Biomedical Research, Institute of Molecular Biology and Biotechnology (FORTH),
GR-451 10, Ioannina, Greece
§Dept. of Materials Science and Engineering, Unit of Medical Technology and Intelligent Information Systems,
University of Ioannina, GR-451 10, Ioannina, Greece
¶Dept. of Electrical and Computer Engineering, Hellenic Mediterranean University,
GR-710 04 Heraklion, Crete, Greece

***Abstract*—Cognitive workload is a critical feature in related psychology, ergonomics, and human factors for understanding performance. However, it still is difficult to describe and thus, to measure it. Since there is no single sensor that can give a full understanding of workload, extended research has been conducted in order to present robust biomarkers. During the last years, machine learning techniques have been used to predict cognitive workload based on various features. Gaze extracted features, such as pupil size, blink activity and saccadic measures, have been used as predictors. The aim of this study is to use gaze extracted features as the only predictors of cognitive workload. Two factors were investigated: time pressure and multi tasking. The findings of this study showed that eye and gaze features are useful indicators of cognitive workload levels, reaching up to 88% accuracy.**

I. Introduction

Cognitive workload can be described as a mental construct that reflects the mental strain resulting from performing a task under specific conditions, coupled with the capability of the operator to respond to those demands [1].

Several studies have focused on the identification of cognitive workload relying solely in eye features for different tasks. Most of them report binary classification results i.e. high and low level of cognitive workload with some of the studies reporting highly accurate results [2], [3], [4]. However, there are only a few reported efforts that focus on multi-class classification (high/medium/low) [5] and in this case the achieved performance is lower.

In the literature, a large variety of eye features have been shown to be useful predictors of cognitive workload. Pupil size seems to be the most useful indicator of cognitive load. However, blink and saccade related features also seem to be correlated with the cognitive workload [6].

The present work involved an experimental study in which participants performed a visual search task together with a secondary demanding working memory task during which, an eye tracking setup was used. At the end of the experimental protocol, the participants filled the NASA-TLX questionnaire [8]. The extracted features from all acquired gaze signals were used as basis for a comparative study between different classification algorithms, including decision trees, discriminant analysis, support vector machine (SVM), k-Nearest Neighbor (kNN) and ensemble learning algorithms, for providing a detailed evaluation of utilizing machine learning to accurately identify between the arousal and valence levels. To our knowledge, this is the first NASA-TLX based workload estimation attempt exploiting solely eye tracking data.

II. Data collection

37 participants (22 females, 15 males) with mean age 29 (SD:7) years were recruited and the mean binocular visual acuity at 80 cm was -0.10±0.07 logMAR. Mean illuminance at cornea when screen was on, was 450 (SD:24) lux.

The study had a 2x2 factorial design, with the two factors being time pressure (with or without) and single vs dual task. The combination of these factors determined four experimental task conditions. Time pressure was imposed asking the participants to complete the task "as fast as they could", while the "no time pressure" task was imposed when the participants were asked to execute the task "with a comfortable pace".

The main task of the study was a visual search task based on a reCAPTCHA-like test, as seen in Fig. 1. A set of images of indoor scenes taken from the free database "Indoor scene recognition" [7] were presented to the participants and they were asked to solve the CAPTCHA-like puzzles. In the dual task, participants were asked to execute an interference task i.e. to perform a backward counting from 1000 by subtracting 4 while executing the main visual search task.

All participants performed 20 trials/images in different conditions (5 trials for each condition/task). Tasks were presented in random order. At the end of each task the participants were

Fig. 1. A sample trial/image of the reCAPTCHA test. Instructions: "Choose the squares in which candles are located."

asked to complete the NASA-TLX questionnaire, a subjective assessment tool that rates perceived workload. The design of the study is shown in Fig. 2.

The reCAPTCHA-like images were presented on a screen at a distance of 80cm from the participant as can be seen at Fig. 3. All measurements were performed with the subjects seated on a chair with their head stabilized by means of a chin and head rest to minimize head movements. Eye tracking measurements were recorded with the Pupil Labs "Pupil Core" gaze tracker (https://pupil-labs.com/products/core/).

From the 20 trials/images the eye and gaze data were processed and analyzed to ensure that the participants did not close their eyes for a duration longer than the average blink or look away from the computer screen for long period of time. If any of the aforementioned cases is true, the relevant data are omitted from the subsequent processing. A total of 740 examples were collected. From these examples, the classes in which the data were split are shown in Table I.

TABLE I
COGNITIVE WORKLOAD CLASSIFICATION CATEGORIES

Cognitive workload score	Classes	Sample size
Mental workload score	high / not high	(186 / 554)
NASA-TLX mean score	high / not high	(186 / 554)
Mental workload score	high / low / medium	(186 / 260 / 294)
NASA-TLX mean score	high / low / medium	(101 / 172 / 467)

The study protocol was approved by the Ethics Committee of FORTH and all participants have signed written consent.

III. METHODOLOGY

A. Parameter extraction and processing

In order to distinguish between the levels of cognitive workload with high efficiency and precision by utilizing only low level eye and gaze data from the eye tracker, it is critical to extract the parameters that can become useful workload indicators. In addition, the features are extracted based on the eye and gaze metrics from the eye tracker. In total, 29 fixations, saccades, blinks and pupil related features are extracted and are shown in Table II. Fixations and saccades are identified based on the I-VT algorithm proposed in [9].

TABLE II
FEATURES EXTRACTED FROM EYE AND GAZE DATA

Fixations	Saccades	Blinks	Pupil
duration*	duration*	frequency	diameter*
total duration	velocity*	duration*	difference from
frequency	frequency		baseline metric
	amplitude*		

* max, min, mean and median values are calculated.

To amend the inequality between the number of data annotated in different classes and consequently to avoid the ineffectiveness of the model to learn the decision boundary, we generated synthetic samples based on the SMOTE oversampling technique [10]. Then, taking into consideration that a) data was normally distributed and b) most machine learning algorithms perform better when numerical input variables are scaled to a standard range [11], we used the MinMax Scaler to scale the features in the range 0-1.

B. Feature selection

After the features were extracted, we built a correlation matrix to study which are highly correlated with each other. Then, in an attempt to derive the most dominant features that could improve the efficiency of our machine learning models, a regularization method was employed and the results obtained were compared with the ANOVA test. Finally, we estimated feature importance using ensemble methods. Alternatively, we performed LASSO regularization analysis, which does both variable selection and regularization to enhance accuracy and interpretability [12].

C. Training and testing

In total, 11 classifiers were examined and tested during the classification procedure which is divided into binary and multi-class. We split the data into training and testing, with the number of the test data being 20% of the total number of examples. The classification algorithms employed are shown in Table III.

TABLE III
CLASSIFICATION ALGORITHMS TESTED

Binary	Multi-class
Gaussian Naive Bayes	
Random Forest	
Linear Support Vector Machine (SVM)	
Ensemble Gradient Boosting	
K-Nearest Neighbor (KNN)	Decision Tree
Linear Perceptron	Ensemble Extra Tree
	Bernoulli Naive Bayes
	Logistic Regression
	Extra Tree

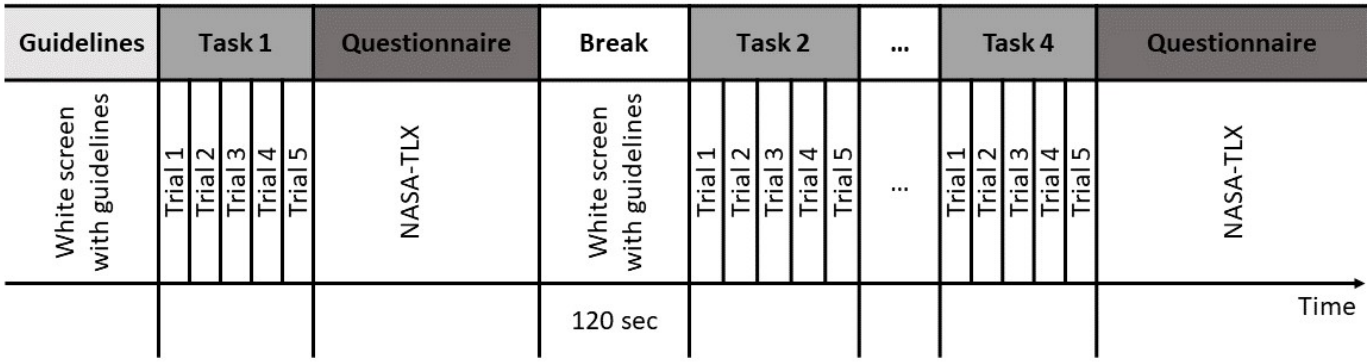

Fig. 2. Design of the study

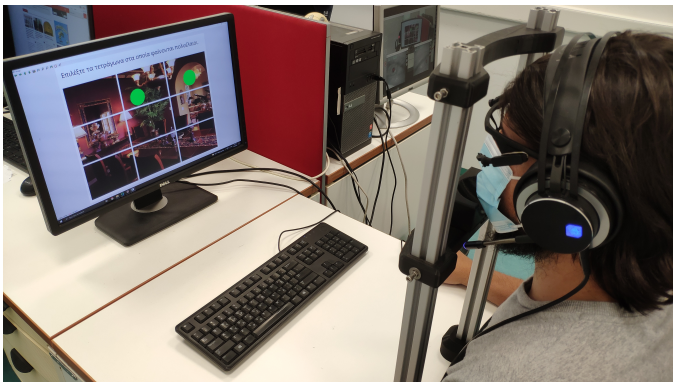

Fig. 3. The experimental setup

D. Hyperparameter tuning

To fine tune the hyperparameters of the proposed model we performed a RandomSearch iterating 1000 times through training data to find the combination of parameters that maximizes the overall performance and accuracy.

E. Model evaluation

The evaluation of the models constructed was performed based on accuracy as well as precision and recall. Combining precision and recall with an armonic mean, we computed the f1-score. In the multi-class cases, these 3 rates are calculated on a per-class basis. Furthermore, we validated the models using a k-fold cross-validation. In addition, for a more comprehensive and graphical representation of our results we plotted the confusion matrices and ROC curves for each fold, thus illustrating how the ability of the classifier changes as its discrimination threshold varies.

IV. Results

From the 37 participants, a total of 740 valid examples were collected. In this section, we illustrate the experimental results of four classification attempts from which, the first two concern the investigation of existence of high cognitive workload (binary approach) while the other two refer to an additional attempt to discriminate cognitive workload among its respective levels (multi-class approach).

The features which were finally extracted were defined as predictors and as response variables the classes presented in Tables IV and V. The training examples were a total of 592 and their respective numbers for each class before the oversampling process are presented in Tables IV and V. The data processing and the classification procedure were processed in Python based programming environment. The models with the higher accuracy were stored and used later for predictions. For the test data, the cognitive workload level prediction rates, recall, precision and f1-score were extracted for the respective machine learning models chosen for each trial. The test data included 148 examples.

Tables IV and V present the results of each classification procedure as well as the response variables, the sample size, the feature selection method, the superior classifier in terms of accuracy, the precision, recall, f1-score and finally the accuracy of the chosen model.

The results of our attempt to predict the presence of mental workload are presented in Table IV. Almost 9 out of 10 examples were classified correctly by the Random Forest classifier, while LASSO analysis was used for the selection of the dominant features. In this binary classification problem, the sensitivity rate of the ”high” instances was found to be 90% and the respective precision rate achieved was 86%. Moreover, the 90% of the positively classified ”not high” mental workload cases were relevant. The f1-score for this classification trial remains above 87% for both classes.

The Random Forest classifier was proven superior (Table IV). The model achieved to correctly predict the existence of ”high” cognitive workload based on the NASA-TLX mean score with 81% accuracy. The features for this procedure were selected with the LASSO analysis. Furthermore, the model managed to predict 84% of positive identifications of ”high” examples that were actually correct, while the respective percentage for the ”not high” examples was 78%. The recall percentages for ”high” and ”not high” examples are 79 and 84%, respectively. Finally, by combining precision and recall metrics we extracted the f1-score which is about 81% for

TABLE IV
COGNITIVE WORKLOAD BINARY CLASSIFICATION RESULTS

Class	Feature selection method	Classifier	Precision (high/not high)	Recall (high/not high)	F1-score (high/not high)	Accuracy
Mental workload	LASSO	Random Forest	0.86/0.90	0.90/0.85	0.88/0.87	0.88
NASA-TLX mean score	LASSO	Random Forest	0.84/0.78	0.79/0.84	0.81/0.81	0.81

TABLE V
COGNITIVE WORKLOAD MULTI-CLASS CLASSIFICATION RESULTS

Class	Feature selection method	Classifier	Precision (high/low/med)	Recall (high/low/med)	F1-score (high/low/med)	Accuracy
Mental workload	ANOVA	Random Forest	0.69/0.65/0.72	0.44/0.77/0.85	0.54/0.71/0.78	0.69
NASA-TLX mean score	ANOVA	Extra Trees	0.87/0.84/0.82	0.98/0.67/0.88	0.92/0.75/0.85	0.84

"high" and "not high" instances.

The last problem is related to the classification of three levels of mental workload; high, medium and low. The Random Forest classifier was once again the most efficient in terms of accuracy reaching up to 69% correct predictions. In more detail, correctly predicting the instances of "medium" mental workload achieved the highest precision, recall and f1-score percentages while the respective scores for the other two mental workload levels remained lower.

Superior results are achieved within the next classification procedure, where we attempted to identify between the three levels of cognitive workload based on the mean score of NASA-TLX test, low, medium and high. In this multi-class problem 84% of the examples were predicted correctly (Table V). The best classification algorithm in terms of accuracy was the Extra Trees and the selection of the dominant features was performed by ANOVA analysis. The precision, recall and f1-score rates of "high" cognitive workload instances were 87, 98 and 92% respectively. In parallel, the Extra Trees achieved satisfactory precision, sensitivity and f1-score for the other two classes.

In summary, the binary classification of mental workload into "high" and "not high" using the Random Forest model achieved the most successful prediction rate 88%. However, when the "medium" class was added to create a multi-class problem, this percentage was reduced by 19%. Finally, regarding the cognitive workload level recognition based on the NASA-TLX score, the success rates of the binary and multi-class problems differ by 3%, with the multi-class identification being more effective. In parallel, the identification of the negative and positive instances of "high" cognitive load level, which demonstrates significant mental effort, was correct in the 98 and 87% of the predicted cases, respectively.

V. CONCLUSIONS

This manuscript presents the results of a study focused on investigating the potential to identify and classify the levels of cognitive workload based on low level eye and gaze features. To this aim, an experimental procedure was designed and performed, for collecting eye and gaze tracking data from participants performing visual search and interference tasks and self-accessing their performance using the NASA-TLX workload index test. From the performed experimental trials certain eye and gaze related identification parameters were extracted and processed, while multiple algorithms were tested for utilizing the ones with the highest success rates for making predictions.

From the results presented in Section IV, the highest success rate was observed during the binary classification attempt for the Random Forest classifier between high and not high mental workload with 88%. However, the inclusion of the "medium" class proved to be challenging leading to a significant decrease in the model's performance. Regarding the cognitive load level estimation based on NASA-TLX score, the binary as well as the multi-class identification tasks provided very promising results reaching up to 84% correct predictions for the multi-class case. These findings provide a potential mechanism for estimating the level of cognitive workload based solely on eye and gaze related features.

Overall these findings are in accordance with findings reported by [2], [3], [4] regarding the binary classification of cognitive workload. However for a more discrete workload level identification, our results go beyond previous reports such as [5], showing the need to continue investigating towards this direction.

VI. FUTURE WORK

Future work is necessary to validate the conclusions drawn from this study. The results must be replicated at a larger scale by adding more participants. Furthermore, it will be important that future research investigate the potential of utilizing deep learning in order to examine their efficiency in the cognitive workload identification problem. We plan also to compare our findings with research works that utilize additional biosignals

for the estimation of cognitive load levels and investigate the necessity and the potential of combining eye and gaze data with other biometrics for increased performance with respect to the computational cost.

ACKNOWLEDGMENT

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] B. Cain. "A review of the mental workload literature," Defence Research and Development Canada Toronto, RTO-TR-HFM-121-Part-II, 2007, Available at: https://apps.dtic.mil/sti/citations/ADA474193

[2] X. Liu, T. Chen, G. Xie and G. Liu, "Contact-Free Cognitive Load Recognition Based on Eye Movement," *Journal of Electrical and Computer Engineering,* November 2016, DOI: https://doi.org/10.1155/2016/1601879

[3] G. Prabhakar, A. Mukhopadhyay, et al, "Cognitive load estimation using ocular parameters," *Automotive, Transportation Engineering*, December 2020, DOI: https://doi.org/10.1016/j.treng.2020.100008

[4] C. Wu, J. Cha, J. Sulek, T. Zhou, C. P. Sundaram, J. Wachs and D. Yu, "Eye-Tracking Metrics Predict Perceived Workload," *Robotic Surgical Skills Training. Hum Factors.* vol. 62, no. 8, pp. 1365-1386, December 2020, doi: 10.1177/0018720819874544.

[5] J. Chen, Q. Zhang, L. Cheng, X. Gao and L. Ding, "A Cognitive Load Assessment Method Considering Individual Differences in Eye Movement Data," in *2019 IEEE 15th International Conference on Control and Automation (ICCA)*, 2019, pp. 295-300, DOI: 10.1109/ICCA.2019.8899595

[6] V. Skaramagkas et al., "Review of eye tracking metrics involved in emotional and cognitive processes," in *IEEE Reviews in Biomedical Engineering*, 2021, doi: 10.1109/RBME.2021.3066072.

[7] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 413-420, DOI: 10.1109/CVPR.2009.5206537

[8] G. Hart and L. E. Staveland, "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research," *Advances in Psychology*, vol. 52, pp. 139-183, 1988, DOI: https://doi.org/10.1016/S0166-4115(08)62386-9

[9] D. D. Salvucci and J. H. Goldberg, Identifying fixations and saccades in eye-tracking protocols," in *ETRA '00: Proceedings of the 2000 symposium on Eye tracking research and applications*, November 2000, pp. 71–78, DOI: https://doi.org/10.1145/355017.355028

[10] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. Philip Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[11] M.M. Ahsan, M.A.P Mahmud, P.K. Saha, K.D. Gupta and Z. Siddique, "Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance," *Technologies*, vol. 9, no. 52, 2021, doi:10.3390/technologies9030052.

[12] R. Muthukrishnan and R. Rohini, "LASSO: A feature selection technique in predictive modeling for machine learning," in *2016 IEEE International Conference on Advances in Computer Applications (ICACA)*, 2016, pp. 18-20.