Cognitive Workload Assessment via Eye Gaze and EEG in an Interactive Multi-Modal Driving Task

Ayca Aygun
Ayca.Aygun@tufts.edu
Department of Computer Science, Tufts University
United States

Boyang Lyu
Boyang.Lyu@tufts.edu
Department of Electrical and Computer Engineering, Tufts University
United States

Thuan Nguyen
Thuan.Nguyen@tufts.edu
Department of Computer Science, Tufts University
United States

Zachary Haga
zachary.haga@tufts.edu
Department of Computer Science, Tufts University
United States

Shuchin Aeron
shuchin@ece.tufts.edu
Department of Electrical and Computer Engineering, Tufts University
United States

Matthias Scheutz
Matthias.Scheutz@tufts.edu
Department of Computer Science, Tufts University
United States

ABSTRACT

Assessing the cognitive workload of human interactants in mixed-initiative teams is a critical capability for autonomous interactive systems to enable adaptations that improve team performance. Yet, it is still unclear, due to diverging evidence, which sensing modality might work best for the determination of human workload. In this paper, we report results from an empirical study that was designed to answer this question by collecting eye gaze and electroencephalogram (EEG) data from human subjects performing an interactive multi-modal driving task. Different levels of cognitive workload were generated by introducing secondary tasks like dialogue, braking events, and tactile stimulation in the course of driving. Our results show that pupil diameter is a more reliable indicator for workload prediction than EEG. And more importantly, none of the five different machine learning models combining the extracted EEG and pupil diameter features were able to show any improvement in workload classification over eye gaze alone, suggesting that eye gaze is a sufficient modality for assessing human cognitive workload in interactive, multi-modal, multi-task settings.

CCS CONCEPTS

• **Human-centered computing** → *HCI design and evaluation methods.*

KEYWORDS

cognitive workload classification, pupillometry, eye gaze, EEG, multi-modality learning, autonomous interactive systems, mixed-initiative teams, artificial agents.

ICMI '22, November 7–11, 2022, Bengaluru, India
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9390-4/22/11.
https://doi.org/10.1145/3536221.3556610

ACM Reference Format:
Ayca Aygun, Boyang Lyu, Thuan Nguyen, Zachary Haga, Shuchin Aeron, and Matthias Scheutz. 2022. Cognitive Workload Assessment via Eye Gaze and EEG in an Interactive Multi-Modal Driving Task. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '22), November 7–11, 2022, Bengaluru, India.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3536221.3556610

1 INTRODUCTION

Cognitive or mental workload, i.e., the extent to which cognitive resources are available and utilized during task execution, is an important determinant of human performance, with a higher workload typically leading to worse performance (e.g., [25]). In the context of mixed-initiative teams, this means that artificial agents should be aware of human cognitive load and not act in ways that would unnecessarily increase it (e.g., through frequent verbal interactions). Various methods have been proposed that would allow artificial agents to determine human cognitive workload during task performance, in particular, electroencephalography (EEG) and human eye gaze [8, 20, 22, 36, 52] both of which have advantages and disadvantages (which we briefly discuss in the next section).

In this paper, we investigate the utility of eye gaze, specifically pupil diameter, and EEG for assessing human workload in an interactive multi-modal driving simulation with four levels of workload and show that eye gaze is the most reliable predictor. Specifically, we experimentally generate different levels of cognitive workload by requiring subjects (drivers) to simultaneously perform multiple tasks such as braking, dialogue interactions, and tactile discrimination in the course of driving. By construction, the cognitive load increases when multiple tasks have to be performed simultaneously. The average percentage change in pupil size (APCPS) [62] is then used as a predictor for the cognitive load. To evaluate the effectiveness of different eye gaze vs. EEG in cognitive workload prediction, we employ five machine learning models: *(a)* k-Nearest Neighbor (k-NN), *(b)* Naive Bayes (NB), *(c)* Random Forest (RF), *(d)* Support-Vector Machines (SVM), and *(e)* Neural Network-based model (NNM) on pupil diameter signal, EEG data, and their combined features. The results show that the combined eye gaze and

EEG data does not significantly improve workload detection over just eye gaze alone.

The main contributions of this paper are summarized as follows:

- We propose a new controlled driving simulation environment that contains multimodal interaction of different physiological sensor modalities including pupillometry and EEG.
- We verify the effectiveness of pupil diameter in assessing different levels of cognitive workload based on the designed driving simulation experiment. Our numerical results show that a notable improvement in cognitive workload classification can be achieved by using pupil diameter as compared to EEG.
- We show that combining the extracted features of EEG and pupil diameter does not improve the accuracy of cognitive workload prediction.

The remainder of this paper is structured as follows. After summarizing related work in Section 2, we briefly introduce the driving simulation environment and provide the details of the experimental setting in Section 3. We then define cognitive workload levels in Section 4, describe the machine learning models in Section 5, and provide the result from training the models in Section 6, followed by a brief conclusion in Section 7.

2 RELATED WORK

EEG is known as a reliable index to detect electrical activity in the brain and is used as a precise measure of cognitive effort [8, 22, 45, 52]. Berka et al. used EEG signals acquired from eighty healthy participants to characterize the correlation between task engagement and cognitive workload while performing learning and memory tasks [8]. In [45], the authors utilized independent component analysis (ICA) to obtain several independent features from EEG and then predict the cognitive workload based on these extracted features. So et al. [52] used short-term frontal EEG signals which were recorded from twenty healthy subjects performing four cognitive and motor tasks to evaluate the dynamic changes of mental workload. Although EEG can be obtained directly from the electrical activity of cortical and subcortical neurons with a very high temporal resolution of milliseconds, it is non-stationary and heavily suffers from undesired noise, such as frequency interference, blinking, motion-related, and sensor-based artifacts [46]. Popular denoising algorithms for EEG signals include wavelet transform-based methods, independent component analysis (ICA)-based techniques, and adaptive filtering [27]. Even though these methods are effective in cleaning particular types of noise, two or more methods are usually required in practice to simultaneously deal with various kinds of noise involved in EEG [27]. However, applying multiple denoising algorithms at the same time not only will increase the complexity of the cleaning process but also may lead to the uncontrolled interaction between these algorithms [47]. In addition, cleaning EEG signals requires the appropriate experiences and sometimes manual processing steps [17].

Eye gaze parameters can also capture the workload fluctuations occurring in a short time interval, leading to the possibility of real-time cognitive workload prediction [2]. Compared to EEG, the human gaze signal is easier to collect in daily living conditions and is less vulnerable to undesired noise and motion artifacts, making it frequently used for workload prediction in various studies such as driving simulation [38, 39]. There have been numerous studies that use pupil diameter as an indicator of cognitive workload levels [6, 9, 15, 39, 41, 42]. Pfleging et al. [42] characterized the relationship between pupil diameter and cognitive effort under several regulated lighting conditions. In [39], Palinko et al. suggested that using human gaze tracking is one of the feasible ways to predict cognitive workload in the course of driving where pupil diameter is a good indicator of cognitive efforts. In [9], Bitkina et al. investigated the performance of a set of eye-tracking metrics such as gaze fixation, pointing, and pupil diameter in predicting driving perceived workload. Pang et al. [41] used multiple eye movement parameters which include fixation duration, blink duration, and pupil diameter to assess different cognitive workload levels by varying the difficulty of web search tasks. There have been relatively few studies that utilize gaze parameters for real-time workload prediction [5, 33]. One study leveraged normalized values of different gaze parameters including pupil diameter, blink duration, and the number of blinks for online classification of cognitive workload [5]. Another study proposed an approach for real-time prediction of mental effort in the context of task complexity adaptation [33].

A few studies investigated combinations of eye gaze and EEG to jointly predict the cognitive workload levels [14, 31, 49]. Khedher et al. [31] collected both eye gaze and brainwave signals of fifteen students during an interaction with a virtual learning environment to identify two groups of learners: students who successfully resolved the tasks and students who did not. Their numerical results indicated that the k-Nearest Neighbor (k-NN) classifier achieves the best accuracy over the six evaluated models with the combination of eye movement and EEG signals. Another study integrated EEG features with pupil diameter to assess the cognitive load [49]. Their design demonstrated that the combination of multiple models has a better performance in classifying different workload levels compared to single models. Even though combining multiple signals may boost-up the accuracy of the trained model, there is no common consensus among researchers about whether combining EEG and pupil diameter signals will provide an advantage in predicting the workload states. For example, Borys et al. [11] studied several combinations of EEG and pupillometry features in arithmetic tasks to point out that models based on eye-tracking features alone achieved higher accuracy in cognitive workload classification than fusion-based models. Another study proposed a method to combine several features obtained from different physiological signal modalities such as EEG, EOG, and human gaze to estimate cognitive workload during simulated remote piloting [14]. The authors claimed that the prediction performance barely decreases with the usage of only EEG compared to the fusion of multiple signal modalities.

It is worth noting that there are a few studies that combine various signal types in addition to EEG and eye gaze. For example, Taamneh et al. [56] investigates the effects of different distraction types such as emotional, cognitive, and sensorimotor on multiple physiological signal modalities including eye gaze, respiration rate, and heart rate during simulated driving.

Finally, we refer the readers to recent surveys on using pupillometry information, EEG signals, or combining both for categorizing the mental workload levels [16, 63].

3 EXPERIMENTAL SETUP

We developed a comprehensive multi-modal interactive experimental paradigm to be able to evaluate the utility of eye gaze vs. EEG for assessing human workload. We decided on a multi-task setting using a driving simulator that required subjects to concentrate on accident-free driving while also engaging in brief dialogue interactions at different times with a confederate. In addition, to increase the cognitive load, subjects also had to perform a tactile detection response task (DRT) at different times. We will next describe the details of the experimental setting, what data we collected and how we processed, followed by data analyses and a discussion of our results.

3.1 Apparatus

We used a medium-fidelity partial-cab driving simulator, which included automatic gear, steering wheel, brake pedal, and accelerator pedal. Five 45-inch liquid crystal displays (LCDs) were used to illustrate the driving environment. The software and hardware equipment were supplied by RTI Health Solutions (Ann Arbor, MI). The simulation consisted of a four-lane highway (each direction included two lanes) with a speed limit of 65 mph. Participants were asked to wear earbuds (Bose QuietComfort 20) to eliminate external noises. During the simulation, participants wore a cylindrical vibrotactile motor on their right collar shoulder which was 14 mm in diameter and 4.5 mm thick for the DRT task and participants were asked to respond to tactile vibrations that happened randomly every 6 to 10 seconds via a response button attached to their right index fingertip. Figure 1 shows a participant from different perspectives while performing the driving task. Finally, from our experiment, multiple physiological signals were collected which are listed as follows:

- EEG: In this study, we recorded EEG signals via eight channels (FC1, FC2, FC5, FC6, CP1, CP2, CP5, and CP6) using 3.14 cm^2 silver/silver chloride electrodes. We used An Enobio (Neuroelectrics, Cambridge, MA, USA) system with a 24-bit resolution to collect the EEG data with a sampling rate of 500 Hz.
- Eye Tracking: In this experiment, we used a Pupil Core (Pupil Labs, Berlin, Germany) eye tracker to assess eye gaze parameters which include a 200 Hz binocular camera and a 120 Hz world camera.

3.2 Participants

The dataset included 80 participants who were recruited from the local community to engage in a single-session study that lasted approximately 120 minutes. The mean age of the participants was 20 with a standard deviation of 3. 46.8% of the participants were identified as female and the rest of them were identified as male. All of the participants were right-handed with a normal or corrected to normal vision and had a valid driver's license. We asked participants to either receive $20 or two hours of research credit for an introductory Psychology course. We also requested participants to complete other driving history and demographics after providing a Declaration of Helsinki consent.

3.3 Design

The driving task included two driving scenarios, one with DRT and one without DRT (non-DRT). Each participant completed two sessions subsequently with a break between them. Half of the participants performed the DRT session in the first part of the experiment besides the rest of them completed the DRT session in the second part of the experiment after the break. Each scenario included a 52.4 km driving simulation and lasted approximately 20 minutes. The first three minutes of each session included only driving without any other event to ensure that the driver was acclimated to the simulation.

Each session (both DRT and non-DRT) included ten braking events where a vehicle appeared 200m in front of the driver. Participants approached the lead vehicle until it was about 75m ahead and then followed it at a fixed distance of 75m for 20 seconds. Six of the ten trials were braking events where the lead vehicle rapidly decelerated for five seconds while its brake lights were activated. At the end of the braking event, the lead vehicle accelerated and moved away from the driver. Four of the ten events were "lure braking events" which are similar to real braking events; however, after 20 seconds, the lead vehicle accelerated away from the driver and did not brake. The order of braking and lure braking events were presented in different orders across participants to eliminate any order effects.

The simulation contained a series of "yes/no" and explanation dialogue interactions. There were 40 questions in total. We asked participants to respond to 20 questions roughly every 30 to 60 seconds during each session. To evaluate the impact of braking events on dialogue performance, we generated different combinations of braking and dialogue events based on their relative timing order. Specifically, we adjusted the time interval between the beginning of the braking event and the completion of the question, i.e., the "stimulus onset asynchrony" (SOA), between -1 and +1 seconds with a step size of 0.5 seconds. SOA values of -1 and -0.5 mean that the braking event takes place 1 and 0.5 seconds after the end of the question, respectively. Similarly, SOA values of 1 and 0.5 indicate that the braking event occurred 1 and 0.5 seconds before the end of the question, respectively. SOA of 0 represents the condition where the braking event started at the same time as the question ends. Each session thus included five SOA events.

4 METHODS

4.1 Workload Evaluation

We generated four cognitive workload levels based on the combination of different events shown in Table 1. In particular, the first level (Level 0) represented the baseline, i.e., the lowest level of cognitive workload that included only the driving task and did not contain any additional events. Workload levels were increased by adding other events such as dialogue, braking, and DRT to generate Level 1, Level 2, and Level 3, respectively. Specifically, Level 2 was created by combining dialogue and braking events (SOA) while Level 3 was generated via combining SOA and DRT events taken from DRT sessions.

Although both "yes/no" and explanation questions were exploited during our experiment, we leveraged only explanation questions to generate higher workload levels under the assumption

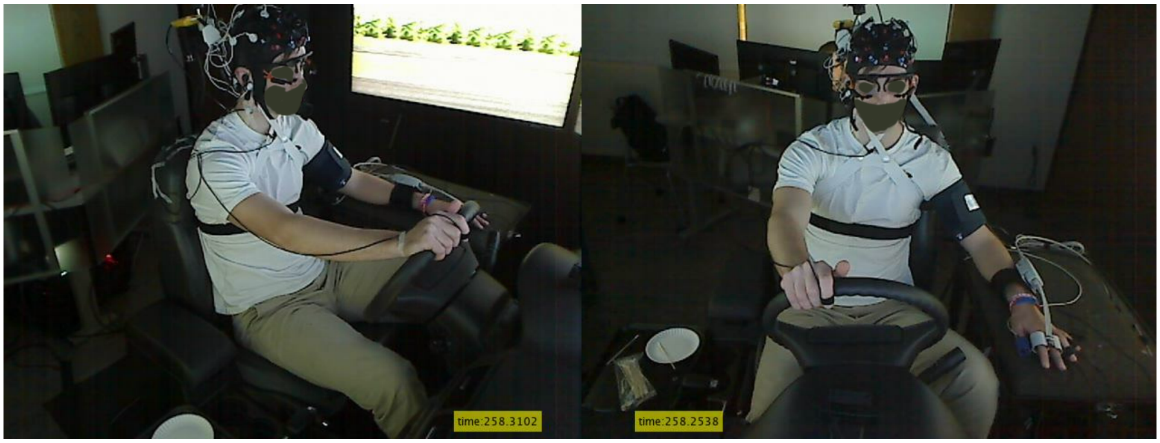

Figure 1: A participant performing the driving task.

Table 1: Cognitive workload levels.

Levels	Dialogue	Braking	DRT
0	-	-	-
1	✓	-	-
2	✓	✓	-
3	✓	✓	✓

that explanation questions should require higher cognitive effort than "yes/no" questions and, therefore, increase cognitive workload. Some examples of explanation questions included "What type of food do you like?", "How often do you drive?", "What is your favorite season", or "What type of movies do you like?".

We note that by construction, the higher the workload level is, the lower the number of samples can be generated. Since there were only 77 SOA events with DRT from the total of 47 subjects, the number of samples in Level 3 was, therefore, 77. To create a balanced dataset, we selected the number of samples in four workload levels equal to the number of samples in Level 3, leading to a balanced dataset with 77 samples in each class. It is also worth noting that Level 0, Level 1, and Level 2 were produced from non-DRT sessions, while Level 3 was produced from DRT sessions. For a fair comparison, the same number of SOA types were used to generate samples for both Level 2 and Level 3. For example, suppose two samples of Level 3 were generated from two SOA and DRT events during the DRT session of a subject with SOA types of -1 and 0.5. In that case, two SOA events were taken from the non-DRT session of the same subject with the same SOA types of -1 and 0.5 to produce two samples of Level 2. Similarly, two dialogue events and two baseline events were taken from the non-DRT session of the same subject to generate two samples of Level 0 and Level 1, respectively.

4.2 Pupillometry

In this study, we used the left pupillometry signal with a sampling frequency of 400 Hz to calculate pupil dilation, assuming that the left and right pupil dilations are synchronous. We utilized three-step pre-processing to eliminate any out-of-band and sensory noise and the blink artifact. In the first step, we applied amplitude thresholding to remove the signal parts lower than 0.8 mm and greater than 10 mm by considering that the values lower than 0.8 mm are potential blink artifacts [50] and the measurable pupil dilation widens up to 10 mm [58]. In the second step, we leveraged linear interpolation to repair the extracted parts [50]. Finally, we used fifth-order Butterworth low-pass filter with a cutoff frequency of 10 Hz to cancel baseline wander [51]. Figure 2 depicts the pre-processing steps of the pupillometry signal.

Due to the range of pupil diameter variations from person to person, we utilized *percentage change in pupil size* (PCPS) to avoid subject-based variations in pupil diameter. We calculated PCPS using the following equation [62]:

$$\mathrm{PCPS} = \frac{\mathrm{CMPD} - \mathrm{BMPD}}{\mathrm{BMPD}} \times 100\%, \tag{1}$$

where CMPD denotes the current measure of pupil diameter and BMPD denotes the baseline measure of diameter which was determined via calculating the mean of a 1-second signal before the stimulus. For each workload level, we obtained 2.5 seconds time windows from pre-processed pupillometry signal and then calculated PCPS values. Next, based on PCPS, we obtained the average PCPS (APCPS) as follows [62]:

$$\mathrm{APCPS} = \frac{1}{N} \sum_{t=1}^{N} \mathrm{PCPS}_t, \tag{2}$$

where PCPS_t denotes the percentage change in pupil size at t^{th} sample, and N is the total number of samples in the time domain.

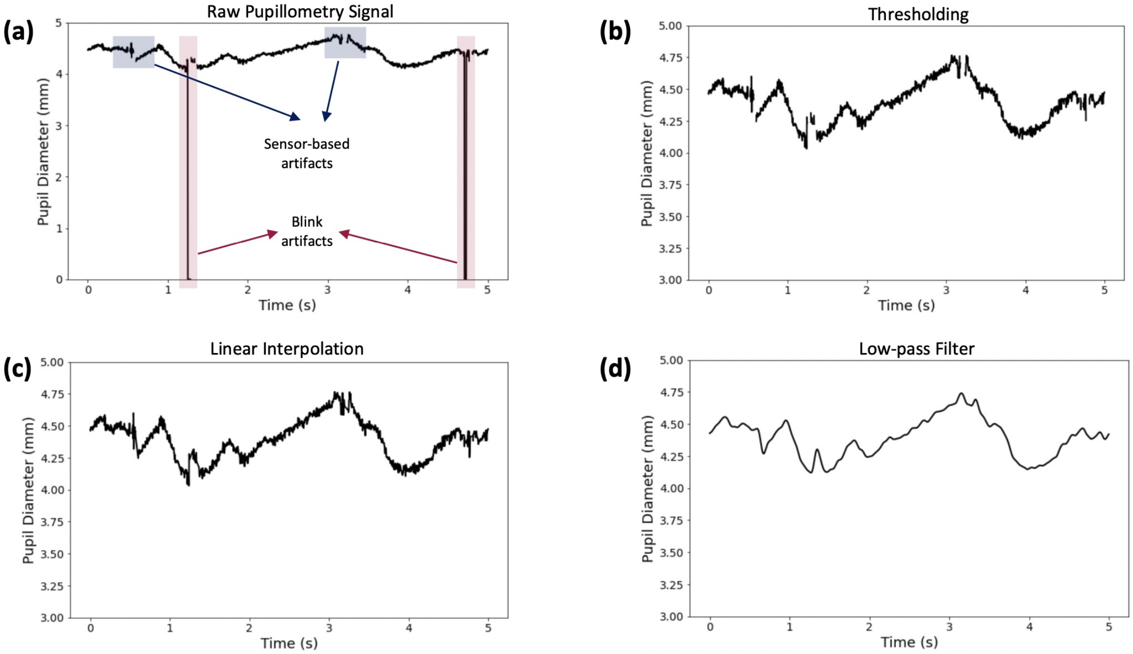

Figure 2: Pupillometry pre-processing steps: (a) raw pupil diameter signal, (b) signal after applying amplitude thresholding, (c) signal after applying linear interpolation, and (d) signal after applying Butterworth low-pass filter.

4.3 Electroencephalography

We recorded EEG signals via eight channels with a sampling rate of 500 Hz and then pre-processed the raw EEG signals to remove the external noise. First, we applied sixth-order Butterworth band-pass filter between 0.1Hz - 32Hz. Second, we leveraged independent component analysis (ICA) to decompose the mixture of signal epochs to its statistically independent components. We manually removed the ICA component which is associated with blink artifacts based on the blink signal. Then, we identified the blink artifact with the instantaneous spikes in the amplitude. Third, we applied Kalman Smoother, a well-known method to estimate the state of dynamic linear structures in the presence of noise [29]. We used the Python library "Tsmoothie" [12] to smooth the EEG signals.

4.4 Feature Extraction

4.4.1 Power Spectral Density of EEG. The Power Spectral Density (PSD) of EEG is one of the most widely used features for EEG signals [3, 21, 24, 44]. Specifically, PSD measures the power distribution of a given signal for each frequency [54]. From the EEG data, we extracted the PSD using the five standard frequency bands: δ (1 to 4Hz), θ (4 to 8Hz), α (8 to 13Hz), β (13 to 30Hz), and γ (30 to 100Hz). Since each EEG sample had the length of 2.5 seconds, we used a periodogram with a 2.5 second non-overlapping rectangular window to estimate the PSD via the MATLAB Signal Processing Toolbox. The periodogram PSD estimator produced the average spectral power over each frequency using Discrete Fourier Transform (DFT) [3, 21]. The average spectral power was then integrated over each EEG frequency band to generate the PSD data. Since there were eight EEG channels and five frequency bands, each PSD sample corresponded to a 40-dimensional vector.

4.4.2 Feature for pupil diameter data. Mean and variance are two commonly extracted features for pupil diameter signals [35, 43, 48]. We compute the mean and variance of each pupil diameter sample processed in Section 4.2. The resulting pupil diameter feature was a two-dimensional vector.

4.4.3 Combination of extracted features of pupil diameter and EEG. To explore the effectiveness of the combination of EEG and pupil diameter data in predicting workload levels, we followed [31, 35, 49, 57] to concatenate the extracted features from EEG and pupil-diameter data, leading to a 42-dimensional feature vector for each sample.

5 MACHINE LEARNING METHODOLOGIES

Based on the data described above, we wanted to explore the performance of different kinds of machine learning techniques for cognitive workload prediction under *(1)* a single-modality setting (i.e., using only the EEG signal or only the pupil diameter signal), compared to *(2)* a multiple-modality setting (i.e., a combination of extracted features from pupil diameter and EEG signals). Since different learning algorithms might have different performances for different types of data, we employed five commonly used machine learning models: k-Nearest Neighbor (k-NN), Naive-Bayes (NB), Random Forest (RF), Support Vector Machines (SVM), and Neural Network-based model (NNM) over six workload classification tasks (which we will discuss in the next section) to make the comparison as far as possible. It is worth noting that these learning models have been widely used for workload classification, for example, k-NN is used in [10, 13, 28, 31], NB is used in [19, 23], RF is used in [28, 40], SVM is used in [4, 18, 37, 52, 60], and NNM is used in [13, 28, 59, 61].

We used the following model training and evaluation procedure. We divided each dataset into two subsets, the training and the testing set with the ratio of 80% and 20%, respectively. For all non-neural network-based models, we utilized five-fold cross-validation [55] for hyper-parameter selection. Specifically, the training set was first split into five equal-size groups. Next, one group was selected as the validation set while the rest four groups were considered as the sub-training set. The model was trained on sub-training and tested on the validation set five times until all groups were selected once as the validation set. The average accuracy on the validation set was used to select the hyper-parameters. After fixing the hyper-parameters, the model was trained from scratch using all training data and tested on the testing set to produce the final classification accuracy. For the neural network-based model (NNM), we kept the setting the same except for not doing the five-fold cross-validation. We performed model selection based on the validation accuracy produced by a fixed validation set and repeated the above procedure five times for all models with different random seeds. Below are the details of the learning models:

5.1 k-Nearest Neighbor

k-Nearest Neighbor is a non-parametric supervised learning method that outputs the label of the tested sample based on the labels of its k-nearest neighbors i.e., the k closest samples to the tested sample in the training set. We used Euclidean distance as the measurement metric and select the value of k in the range of $[1, 30]$.

5.2 Naive Bayes

The Naive Bayes algorithm is a supervised learning method that is based on the well-known Bayes' theorem together with a "naive" assumption such that every pair of samples given the label are independent. There are five common models for the Naive Bayes algorithm: *(a)* Gaussian Naive Bayes, *(b)* Multinomial Naive Bayes, *(c)* Complement Naive Bayes, *(d)* Bernoulli Naive Bayes, and *(e)* Categorical Naive Bayes. These models are different in the way the conditional distribution between the data and its label is modeled. For example, in Gaussian Naive Bayes, one assumes that the conditional distribution of the sample given its label is Gaussian. Here, we considered the types of distribution as the tunable hyper-parameters.

5.3 Random Forest

Random Forest is a supervised algorithm that aggregates multiple decision trees trained on different samples and takes their majority vote for prediction. The number of decision trees (or the number of estimators) was selected in the range of $[1, 50]$.

5.4 Support Vector Machine

Support Vector Machine is a supervised learning algorithm that maps training examples from the input space into the feature space to maximize the width of the gap between the categories using kernel functions. We considered the types of kernel function as our tunable hyper-parameters. Specifically, we selected the model among four kernel functions: linear, polynomial, radial basis function (RBF), and sigmoid.

5.5 Neural Network-based Models (NNM)

Deep Neural Networks have been proved to be effective in extracting task-related features and thus have been widely used in brain-computer interface field [13, 28, 34, 59, 61]. Here, we adopt EEGNet [34] for EEG data and use Multi-layer Perceptron (MLP) neural network for the remaining types of data. As recommended in [34], the length of the temporal convolution of the first layer in EEGNet is set to 250 which is exactly half of the sampling rate of EEG signals (500 samples per second). The MLP model for pupil diameter data is composed of three fully connected layers followed by a linear layer with the output dimension as $512 \longrightarrow 256 \longrightarrow 128 \longrightarrow n$, where $n \in \{2, 3, 4\}$ denotes the number of label classes i.e., the number of workload levels. Rectified linear unit (ReLU) [1] was used as the activation function. To stabilize the training process and avoid overfitting, batch normalization and dropout were applied to each fully connected layer. The remaining two kinds of data (PSD and pupil diameter feature) shared the same model structure with pupil diameter data except for the output dimension of each layer. Particularly, the MLP models for PSD data had the output dimension as $40 \longrightarrow 32 \longrightarrow 16 \longrightarrow n$ for each layer and the MLP model for pupil diameter feature had the output dimension as $32 \longrightarrow 32 \longrightarrow 16 \longrightarrow n$ for each layer. All models were trained with full size of data in one batch for 300 epochs using the Adam optimizer [32] with the learning rate set as 5×10^{-4} for EEGNet and 10^{-4} for other models.

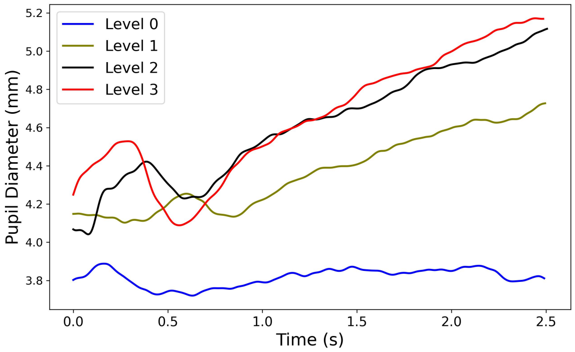

Figure 3: Pupil diameter variations of one participant for different cognitive workload levels within 2.5 seconds time interval.

6 RESULTS AND DISCUSSION

We first investigated the changes in pupil diameter (PCPS) and generated statistical results of PCPS and average PCPS (APCPS) for different workload conditions. Then we reported the classification accuracies of well-known learning methodologies for predicting different workload levels based on the single-modality or multiple-modality settings.

6.1 Statistical Analysis of PCPS

Figure 3 shows pupil diameter variations of one participant for four cognitive workload levels within a 2.5 seconds time interval. The starting point represents the onset of baseline, dialogue, SOA, and SOA+DRT events for Level 0, Level 1, Level 2, and Level 3, respectively. The results demonstrate that the pupil diameter remains stable during the baseline while distinctive patterns are observed for higher workload levels. However, the fluctuations in pupil diameter have analogous patterns for Level 2 and Level 3.

Next, we calculated the mean APCPS values on all events for each workload level and compared the outcomes to evaluate the variation of mean APCPS for various workload levels. To verify a direct correlation between mental effort and APCPS, we performed a one-way ANOVA test [53]. Table 2 and Figure 4 illustrate the mean, standard deviation, and standard error of APCPS values related to four cognitive workload levels, showing that the pupillary response rises as mental effort increases. The results demonstrate that mean APCPS has the lowest value for baseline and increases gradually with the increased cognitive workload. However, the values associated with Level 2 and Level 3 do not indicate a marked difference.

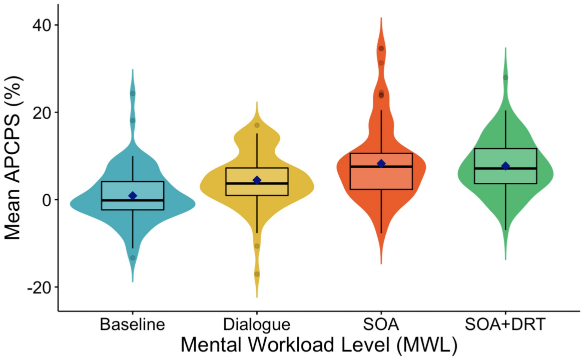

Figure 4: Violin plot of mean APCPS over all events for different cognitive workload levels.

We further run Tukey's "honestly significant difference" (HSD) multiple pairwise comparison test to investigate the differences between the four workload levels [26]. Table 3 illustrates the difference in means, the minimum and maximum confidence levels, and the adjusted p-values for all pairs of workload levels at a significance level of .95. Moreover, we performed Benjamini & Hochberg test [7] to decrease the false discovery rate and obtained p-values, as shown in Table 4. The results of both tests indicate that there is a statistically notable difference between all pairs of workload levels ($p < .05$) except between Level 2 and Level 3 ($p > .05$). The transition between Level 0 and Level 1 was achieved by performing dialogue tasks that included multiple explanation questions. Responding to different questions while completing the driving task requires higher comprehension and cognitive engagement, which was reflected by a remarkable change in pupil dilation between Level 0 and Level 1. Level 2 was generated by adding a braking event to Level 1. There were six braking events and four lure braking events during the entire experimental run which is reasonably rare. Thus, an additional braking event to both dialogue and driving tasks increased workload and caused increased pupil dilation. On the other hand, there is a negligible change in pupil diameter values between Level 2 and Level 3. The possible reason for this insignificant difference is that DRT events occurred frequently every 6 to 10 seconds during the experiment and thus might be less likely to significantly increase cognitive workload in the short term.

6.2 Statistical Analysis of EEG

We also investigated the performance of EEG signal on workload estimation by performing ANOVA Tukey HSD multiple pairwise tests on five different frequency bands generated from PSD of EEG which are delta, theta, alpha, beta, and gamma. The p-values corresponding to each pair of workload levels are illustrated in Table 5. The results indicate that within five EEG frequency bands, just alpha and beta waves are efficient in differentiating workload level pairs 0-2 and 0-3. However, none of the five frequency bands are capable of classifying workload level pairs 0-1, 1-2, 1-3, and 2-3. Moreover, we observed that delta, theta, and gamma waves are not able to distinguish any pairs of workload levels. These results demonstrated that none of the frequency bands is competent for the classification of all pairs of workload levels.

6.3 Classification Performance

As previously mentioned in Section 5, we employ five different machine learning models in this section to evaluate the mental workload classification accuracy under both single-modality and multi-modality settings.

6.3.1 Single-modality learning. Under the single-modality learning framework, we report the averaged accuracy of two-level, three-level, and four-level cognitive workload classification tasks and their standard deviation after repeating all experiments five times. For two-level classification, we report the performance of three tasks: *(a)* the "0-1" task i.e., differentiating between Level 0 and Level 1; *(b)* the "0-2" task i.e., distinguishing between Level 0 and Level 2; and *(c)* the "0-3" task i.e., distinguishing between Level 0 and Level 3. For three-level classification, we report the performance of two tasks: *(a)* the "0-1-2" task i.e., differentiating between Level 0, Level 1, and Level 2; and *(b)* the "0-1-3" task i.e., distinguishing between Level 0, Level 1, and Level 3. Finally, the classification accuracy of distinguishing between four workload levels, denoted as the "0-1-2-3" task, is also reported.

The performance of the two-level classification tasks is shown in Table 6. As seen, with the corresponding highest accuracy as 72.50 ∓ 12.38 for task "0-1", 75.00 ∓ 4.42 for task "0-2", and 75.00 ∓ 5.76 for task "0-3", one can infer that: *(a)* there is no substantial difference in difficulty levels between task "0-2" and task "0-3", and *(b)* task "0-2" and task "0-3" are a bit easier than task "0-1". Since the shared Level 0 between these three tasks indicates the baseline, this observation implies that *(a)* there is a negligible difference in the amount of workload between Level 2 and Level 3, and *(b)* the amount of workload contained in Level 2 and Level 3 is a bit heavier compared to Level 1. Indeed, this conclusion agrees with our observation from Figure 3, Figure 4, and Table 2 in Section 6.1.

Next, from the similarity between the classification accuracy of "0-1-2" and "0-1-3" tasks, as shown in Table 7, one can conclude that

Table 2: Statistical results of mean, standard deviation, and standard error APCPS for different workload levels over all events.

Workload Level	Mean APCPS	Std	SE
Level 0	0.88	6.06	0.69
Level 1	4.45	6.21	0.71
Level 2	8.25	8.69	0.99
Level 3	7.69	6.30	0.72

Table 3: Results of Tukey HSD multiple pairwise test for different pairs of cognitive workload levels.

Workload Pairs	Diff. in Means	Lower Value	Upper Value	P-value
Level 1 - Level 0	3.563	0.690	6.436	0.008
Level 2 - Level 0	7.368	4.495	10.241	9.5×10^{-10}
Level 3 - Level 0	6.809	3.936	9.682	1.7×10^{-8}
Level 2 - Level 1	3.804	0.931	6.677	0.003
Level 3 - Level 1	3.245	0.372	6.118	0.019
Level 3 - Level 2	-0.559	-3.432	2.313	0.958

Table 4: Adjusted p-values obtained from Benjamini & Hochberg test for different pairs of workload levels.

	Level 0	Level 1	Level 2
Level 1	8.5×10^{-4}	-	-
Level 2	3.1×10^{-8}	0.002	-
Level 3	1.1×10^{-9}	0.002	0.648

Table 5: p-values from Tukey HSD multiple pairwise test performed on five frequency bands obtained from PSD of EEG for different pairs of cognitive workload levels.

Workload Level	Delta	Theta	Alpha	Beta	Gamma
Level 1-Level 0	0.23	0.47	0.14	0.18	0.30
Level 2-Level 0	0.97	0.99	$\mathbf{3 \times 10^{-3}}$	**0.03**	0.12
Level 3-Level 0	0.36	0.92	$\mathbf{4 \times 10^{-4}}$	**0.01**	0.13
Level 2-Level 1	0.49	0.58	0.59	0.89	0.96
Level 3-Level 1	0.99	0.85	0.26	0.70	0.98
Level 3-Level 2	0.65	0.96	0.94	0.98	0.99

there is not much difference in the amount of workload contained in Level 2 and Level 3. Again, this agrees with the intuition from Figure 3, Figure 4, and Table 2 in Section 6.1, where it is practically impossible to distinguish the pupil diameter changes between these two levels.

The performance of the four-level "0-1-2-3" classification task is shown in Table 8. As seen, the accuracy of "0-1-2-3" task is significantly lower than other classification tasks. This comes from the negligible difference between the amount of workload contained in levels 2 and 3. Specifically, by construction, level 3 is generated from level 2 by adding DRT events that occur frequently and do not cause a significant increase in cognitive workload in a short period of time. Therefore, it is generally challenging to distinguish between level 2 and level 3, leading to the highest prediction accuracy of the "0-1-2-3" task being only 43.44 ∓ 6.80 achieved by using pupil diameter signal and the NNM method.

Furthermore, from Table 6, 7 and Table 8, the pupil diameter signal outperforms EEG and its extracted feature for all tasks regardless of the machine learning methods. We thus conclude that pupil-diameter data is more suitable for cognitive workload prediction under the current driving simulation setting.

In addition, one can observe that the classification performance based on extracted signals is comparable to their original signals, implying that most of the important information has remained after the feature extraction process. This verification is necessary since the extracted features will be combined as the input for our multi-modality learning in Section 6.3.2.

Finally, it is worth noting that we also applied multivariate long short term memory fully convolutional network (MLSTM-FCN) [30] models for both EEG and pupil diameter signals. For six tasks: "0-1", "0-2", "0-3", "0-1-2", "0-1-3", and "0-1-2-3", the accuracies of EEG are 53.25 ∓ 2.71, 55.65 ∓ 4.44, 54.70 ∓ 9.25, 35.28 ∓ 3.48,

Table 6: Two-level tasks "0-1", "0-2" and "0-3" classification accuracy.

Task	Signals	*k*-NN	NB	RF	SVM	NNM
0-1	Pupil diameter	65.32 ∓ 1.39	68.75 ∓ 9.88	70.31 ∓ 12.00	67.74 ∓ 3.95	**72.50** ∓ 12.38
	EEG	53.96 ∓ 4.19	59.37 ∓ 5.41	46.88 ∓ 6.99	50.81 ∓ 6.03	52.50 ∓ 4.64
	Pupil diameter features	67.74 ∓ 6.03	70.31 ∓ 2.70	60.94 ∓ 6.81	68.54 ∓ 4.19	66.25 ∓ 6.78
	EEG features	50.11 ∓ 5.81	62.50 ∓ 7.65	55.23 ∓ 6.02	56.45 ∓ 3.61	53.12 ∓ 4.42
0-2	Pupil diameter	70.97 ∓ 4.19	65.63 ∓ 10.36	**75.00** ∓ 4.42	71.77 ∓ 4.19	69.38 ∓ 8.67
	EEG	59.68 ∓ 8.06	60.94 ∓ 5.18	60.94 ∓ 6.81	55.65 ∓ 3.51	50.62 ∓ 6.78
	Pupil diameter features	66.12 ∓ 3.61	68.75 ∓ 10.83	59.81 ∓ 8.11	72.58 ∓ 3.61	62.50 ∓ 6.63
	EEG features	53.23 ∓ 5.58	56.25 ∓ 9.88	50.00 ∓ 4.42	52.42 ∓ 2.67	51.87 ∓ 2.79
0-3	Pupil diameter	73.38 ∓ 2.67	67.19 ∓ 12.79	71.88 ∓ 9.38	**75.00** ∓ 5.76	71.25 ∓ 5.61
	EEG	55.64 ∓ 6.98	57.26 ∓ 5.18	55.31 ∓ 6.81	53.25 ∓ 5.34	56.25 ∓ 8.90
	Pupil diameter features	70.35 ∓ 6.65	70.31 ∓ 6.81	67.19 ∓ 9.24	67.74 ∓ 2.28	69.37 ∓ 5.44
	EEG features	57.26 ∓ 8.34	56.25 ∓ 6.25	56.25 ∓ 15.31	53.58 ∓ 2.67	54.50 ∓ 9.25

Table 7: Three-level tasks "0-1-2" and "0-1-3" classification accuracy.

Task	Signals	*k*-NN	NB	RF	SVM	NNM
0-1-2	Pupil diameter	55.85 ∓ 2.76	50.00 ∓ 8.83	55.21 ∓ 5.41	53.19 ∓ 6.56	**56.67** ∓ 3.42
	EEG	35.63 ∓ 3.14	37.50 ∓ 4.16	31.25 ∓ 9.54	35.10 ∓ 2.38	35.83 ∓ 8.25
	Pupil diameter features	51.59 ∓ 4.08	48.72 ∓ 3.45	47.38 ∓ 4.54	48.40 ∓ 1.76	41.67 ∓ 6.07
	EEG features	33.15 ∓ 7.11	43.48 ∓ 5.33	32.61 ∓ 8.96	31.52 ∓ 3.26	37.92 ∓ 5.78
0-1-3	Pupil diameter	52.12 ∓ 4.38	49.83 ∓ 7.80	48.96 ∓ 9.02	55.31 ∓ 2.60	**57.92** ∓ 6.81
	EEG	30.85 ∓ 3.83	39.58 ∓ 8.58	38.54 ∓ 10.36	32.98 ∓ 5.72	31.66 ∓ 5.19
	Pupil diameter features	52.12 ∓ 10.91	48.95 ∓ 8.00	52.08 ∓ 16.79	48.40 ∓ 4.85	39.59 ∓ 6.42
	EEG features	33.69 ∓ 3.92	40.21 ∓ 4.74	31.52 ∓ 6.43	35.86 ∓ 1.88	33.33 ∓ 9.66

Table 8: Four-level task "0-1-2-3" classification accuracy.

Task	Signals	*k*-NN	NB	RF	SVM	NNM
0-1-2-3	Pupil diameter	37.90 ∓ 2.41	39.51 ∓ 4.19	35.48 ∓ 5.59	39.52 ∓ 1.80	**43.44** ∓ 6.80
	EEG	25.40 ∓ 1.33	30.64 ∓ 6.65	25.38 ∓ 8.64	27.42 ∓ 1.61	30.00 ∓ 3.89
	Pupil diameter features	38.71 ∓ 1.98	38.71 ∓ 2.28	32.26 ∓ 3.95	37.90 ∓ 3.33	33.44 ∓ 6.01
	EEG features	27.05 ∓ 7.55	31.45 ∓ 1.40	27.42 ∓ 9.54	26.64 ∓ 1.36	26.25 ∓ 3.39

37.10 ∓ 2.91, 33.05 ∓ 3.58 while the accuracies of pupil diameter are 65.75 ∓ 7.33, 69.10 ∓ 6.81, 67.19 ∓ 7.26, 52.50 ∓ 4.85, 54.22 ∓ 1.89, 36.91 ∓ 4.41, respectively. As seen, the MLSTM-FCN models always achieve lower accuracies than EEGNet and MLP over all six learning tasks. Therefore, we decided not to provide the numerical results of MLSTM-FCN in the main tables but mention here to provide additional information for the readers.

6.3.2 Multi-modality learning. For multi-modality learning, two approaches are considered in this paper: (1) feature-level fusion approach i.e., combining the extracted features of pupil diameter and EEG as a single input of NNM, and (b) intermediate-level fusion approach i.e., using two MLP networks to separately learn the useful features from EEG and pupil-diameter signals and then feed the outputs of these two networks into a third MLP network to output the final labels.

The performances of multi-modality learning for six workload classification tasks are indicated in Table 9. Compared to single-modality learning, it seems that multi-modality learning does not improve the accuracy of mental workload prediction. Specifically, the highest accuracy (over all tested models) and its standard deviation of pupil diameter (the most left bar, blue color), pupil diameter features (the second left bar, red color), EEG (middle bar, yellow color), EEG features (the second right bar, purple color), and the combined features of pupil diameter and EEG (the most right bar, green color) for six workload classification tasks are illustrated via a bar-graph in Figure 5. As seen, the performance of using multi-modality learning is comparable to or slightly better than using pupil diameter features, EEG, and EEG features but significantly lower than using pupil diameter. This leads to our conclusion that combining the extracted feature of EEG and pupil diameter does not improve the quality of workload prediction. It is worth noting

Table 9: Multi-modality classification accuracy.

Signals	Fusion approaches	0-1	0-2	0-3	0-1-2	0-1-3	0-1-2-3
Extracted feature of pupil diameter + EEG	Feature-level	61.25 ∓ 8.44	71.25 ∓ 6.78	72.51 ∓ 8.22	45.42 ∓ 7.43	41.67 ∓ 5.17	33.13 ∓ 8.00
Time-series of pupil diameter + EEG	Intermediate-level	59.38 ∓ 5.59	63.05 ∓ 3.47	61.25 ∓ 1.53	41.25 ∓ 1.56	35.42 ∓ 2.28	30.94 ∓ 1.57

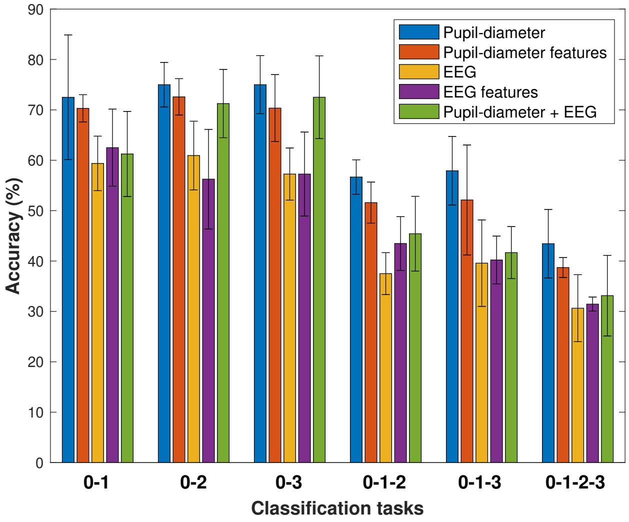

Figure 5: The highest accuracy over all tested methods (in percentage) and its standard deviation using pupil diameter, pupil diameter features, EEG, EEG features, and the combined features of pupil diameter and EEG over six workload classification tasks.

that our conclusion agrees with the previous results observed in [11].

7 CONCLUSIONS

The goal of our empirical study and machine learning efforts was to investigate the potential of eye gaze and EEG, alone or combined, for assessing human workload in a multi-modal interactive driving task. We found out that pupil dilation is an effective method to distinguish different levels of cognitive workload that we generated by adding dialogue, braking, and tactile tasks to the participants' main driving task. Specifically, our analyses showed that the percentage change in pupil size (PCPS) and the average-PCPS (APCPS) are practical tools for differentiating multiple levels of cognitive workload. We also found out that compared to EEG, pupil diameter provides better workload classification and, most importantly, combining the extracted features of EEG and pupil diameter for jointly assessing cognitive workload does not improve the overall prediction accuracy.

Our findings are important for future efforts in online cognitive workload detection in human-machine interaction settings because they support using only eye gaze data which is easier to collect and faster to process. Moreover, eye trackers can be easily applied and worn in many task settings without limiting the mobility of the subject and without introducing motion artifacts like EEG.

8 LIMITATIONS AND FUTURE WORK

Our work has some limitations. First, the proposed method provides an offline assessment of cognitive workload. In the future, we will extend the current setting to real time workload estimation. Second, while this work only uses pupil dilation as a gaze parameter, combining pupil dilation with other aspects of the human eye gaze such as the number of fixations, the fixation duration, the blink rate, and the saccadic intrusion could potentially provide additional information to further improve the workload prediction accuracy. In the future, we will investigate the performance of additional human gaze parameters on cognitive workload estimation.

ACKNOWLEDGMENTS

This project was in part supported by AFOSR grant #FA9550-18-1-0465.

REFERENCES

[1] Abien Fred Agarap. 2018. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375* (2018).

[2] Ulf Ahlstrom and Ferne J Friedman-Berg. 2006. Using eye movement activity as a correlate of cognitive workload. *International journal of industrial ergonomics* 36, 7 (2006), 623–636.

[3] Abeer Al-Nafjan, Manar Hosny, Areej Al-Wabil, and Yousef Al-Ohali. 2017. Classification of human emotions from electroencephalogram (EEG) signal using deep neural network. *Int. J. Adv. Comput. Sci. Appl* 8, 9 (2017), 419–425.

[4] Mohammad A Almogbel, Anh H Dang, and Wataru Kameyama. 2019. Cognitive workload detection from raw EEG-signals of vehicle driver using deep learning. In *2019 21st International Conference on Advanced Communication Technology (ICACT)*. IEEE, 1–6.

[5] Tobias Appel, Christian Scharinger, Peter Gerjets, and Enkelejda Kasneci. 2018. Cross-subject workload classification using pupil-related measures. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*. 1–8.

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

[7] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* 57, 1 (1995), 289–300.

[8] Chris Berka, Daniel J Levendowski, Michelle N Lumicao, Alan Yau, Gene Davis, Vladimir T Zivkovic, Richard E Olmstead, Patrice D Tremoulet, and Patrick L Craven. 2007. EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks. *Aviation, space, and environmental medicine* 78, 5 (2007), B231–B244.

[9] Olga Vl Bitkina, Jaehyun Park, and Hyun K Kim. 2021. The ability of eye-tracking metrics to classify and predict the perceived driving workload. *International Journal of Industrial Ergonomics* 86 (2021), 103193.

[10] Justin A Blanco, Michael K Johnson, Kyle J Jaquess, Hyuk Oh, Li-Chuan Lo, Rodolphe J Gentili, and Bradley D Hatfield. 2016. Quantifying cognitive workload in simulated flight using passive, dry EEG measurements. *IEEE Transactions on Cognitive and Developmental Systems* 10, 2 (2016), 373–383.

[11] Magdalena Borys, Małgorzata Plechawska-Wójcik, Martyna Wawrzyk, and Kinga Wesołowska. 2017. Classifying cognitive workload using eye activity and EEG features in arithmetic tasks. In *International conference on information and software technologies*. Springer, 90–105.

[12] Marco Cerliani. 2021. Tsmoothie. https://github.com/cerlymarco/tsmoothie.

[13] Baljeet Singh Cheema, Shabnam Samima, Monalisa Sarma, and Debasis Samanta. 2018. Mental workload estimation from EEG signals using machine learning algorithms. In *International Conference on Engineering Psychology and Cognitive Ergonomics*. Springer, 265–284.

[14] James C Christensen, Justin R Estepp, Glenn F Wilson, and Christopher A Russell. 2012. The effects of day-to-day variability of physiological data on operator functional state classification. *NeuroImage* 59, 1 (2012), 57–63.

[15] Souvik Das, Kintada Prudhvi, and J Maiti. 2022. Assessing Mental Workload Using Eye Tracking Technology and Deep Learning Models. *Handbook of Intelligent Computing and Optimization for Sustainable Development* (2022), 1–11.

[16] Essam Debie, Raul Fernandez Rojas, Justin Fidock, Michael Barlow, Kathryn Kasmarik, Sreenatha Anavatti, Matt Garratt, and Hussein A Abbass. 2019. Multimodal fusion for objective assessment of cognitive workload: a review. *IEEE transactions on cybernetics* 51, 3 (2019), 1542–1555.

[17] Arnaud Delorme and Scott Makeig. 2004. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of neuroscience methods* 134, 1 (2004), 9–21.

[18] Georgios N Dimitrakopoulos, Ioannis Kakkos, Zhongxiang Dai, Julian Lim, Joshua J deSouza, Anastasios Bezerianos, and Yu Sun. 2017. Task-independent mental workload classification based upon common multiband EEG cortical connectivity. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 25, 11 (2017), 1940–1949.

[19] Aruna Duraisingam, Ramaswamy Palaniappan, and Samraj Andrews. 2017. Cognitive task difficulty analysis using EEG and data mining. In *2017 Conference on Emerging Devices and Smart Systems (ICEDSS)*. IEEE, 52–57.

[20] Tjerk de Greef, Harmen Lafeber, Herre van Oostendorp, and Jasper Lindenberg. 2009. Eye movement as indicators of mental workload to trigger adaptive automation. In *International Conference on Foundations of Augmented Cognition*. Springer, 219–228.

[21] N Hamzah, Haryanti Norhazman, Norliza Zaini, and Maizura Sani. 2016. Classification of EEG signals based on different motor movement using multi-layer Perceptron artificial neural network. *J Biol Sci* 16, 7 (2016), 265–271.

[22] Jamison Heard, Caroline E Harriott, and Julie A Adams. 2018. A survey of workload assessment algorithms. *IEEE Transactions on Human-Machine Systems* 48, 5 (2018), 434–451.

[23] Ryan M Hope, Ziheng Wang, Zuoguan Wang, Qiang Ji, and Wayne D Gray. 2011. Workload classification across subjects using EEG. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 55. SAGE Publications Sage CA: Los Angeles, CA, 202–206.

[24] Md Farhad Hossain, Hamwira Yaacob, and Azlin Nordin. 2021. Development of Unified Neuro-Affective Classification Tool (UNACT). In *IOP Conference Series: Materials Science and Engineering*, Vol. 1077. IOP Publishing, 012031.

[25] Zachary L Howard, Reilly Innes, Ami Eidels, and Shayne Loft. 2021. Using Past and Present Indicators of Human Workload to Explain Variance in Human Performance. *Psychonomic Bulletin & Review* 28, 6 (2021), 1923–1932.

[26] James Jaccard, Michael A Becker, and Gregory Wood. 1984. Pairwise multiple comparison procedures: A review. *Psychological Bulletin* 96, 3 (1984), 589.

[27] Xiao Jiang, Gui-Bin Bian, and Zean Tian. 2019. Removal of artifacts from EEG signals: a review. *Sensors* 19, 5 (2019), 987.

[28] Monika Kaczorowska, Małgorzata Plechawska-Wójcik, and Mikhail Tokovarov. 2021. Interpretable machine learning models for three-way classification of cognitive workload levels for eye-tracking features. *Brain sciences* 11, 2 (2021), 210.

[29] Rudolph Emil Kalman. 1960. A new approach to linear filtering and prediction problems. (1960).

[30] Fazle Karim, Somshubra Majumdar, Houshang Darabi, and Samuel Harford. 2019. Multivariate LSTM-FCNs for time series classification. *Neural Networks* 116 (2019), 237–245.

[31] Asma Ben Khedher, Imène Jraidi, and Claude Frasson. 2019. Predicting learners’ performance using EEG and eye tracking features. In *The Thirty-Second International Flairs Conference*.

[32] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[33] Thomas Kosch, Mariam Hassib, Daniel Buschek, and Albrecht Schmidt. 2018. Look into my eyes: using pupil dilation to estimate mental workload for task complexity adaptation. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–6.

[34] Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. 2018. EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of neural engineering* 15, 5 (2018), 056013.

[35] Jesus L Lobo, Javier Del Ser, Flavia De Simone, Roberta Presta, Simona Collina, and Zdenek Moravek. 2016. Cognitive workload classification using eye-tracking and EEG data. In *Proceedings of the International Conference on Human-Computer Interaction in Aerospace*. 1–8.

[36] James G May, Robert S Kennedy, Mary C Williams, William P Dunlap, and Julie R Brannan. 1990. Eye movement indices of mental workload. *Acta psychologica* 75, 1 (1990), 75–89.

[37] Moona Mazher, Azrina Abd Aziz, Aamir Saeed Malik, and Hafeez Ullah Amin. 2017. An EEG-based cognitive load assessment in multimedia learning using feature extraction and partial directed coherence. *IEEE Access* 5 (2017), 14819–14829.

[38] Oskar Palinko and Andrew L Kun. 2012. Exploring the effects of visual cognitive load and illumination on pupil diameter in driving simulators. In *Proceedings of the Symposium on Eye Tracking Research and Applications*. 413–416.

[39] Oskar Palinko, Andrew L Kun, Alexander Shyrokov, and Peter Heeman. 2010. Estimating cognitive load using remote eye tracking in a driving simulator. In *Proceedings of the 2010 symposium on eye-tracking research & applications*. 141–144.

[40] Vishal Pandey, Dhirendra Kumar Choudhary, Vinita Verma, Greeshma Sharma, Ram Singh, and Sushil Chandra. 2020. Mental Workload Estimation Using EEG. In *2020 Fifth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)*. IEEE, 83–86.

[41] Liping Pang, Yurong Fan, Ye Deng, Xin Wang, and Tianbo Wang. 2020. Mental Workload Classification By Eye Movements In Visual Search Tasks. In *2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. IEEE, 29–33.

[42] Bastian Pfleging, Drea K Fekety, Albrecht Schmidt, and Andrew L Kun. 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*. 5776–5788.

[43] Małgorzata Plechawska-Wójcik and Magdalena Borys. 2016. An analysis of EEG signal combined with pupillary response in the dynamics of human cognitive processing. In *2016 9th International Conference on Human System Interactions (HSI)*. IEEE, 378–385.

[44] Xuemei Qin, Yunfei Zheng, and Badong Chen. 2019. Extract EEG Features by Combining Power Spectral Density and Correntropy Spectral Density. In *2019 Chinese Automation Congress (CAC)*. IEEE, 2455–2459.

[45] Hongquan Qu, Yiping Shan, Yuzhe Liu, Liping Pang, Zhanli Fan, Jie Zhang, and Xiaoru Wanyan. 2020. Mental workload classification method based on EEG independent component features. *Applied Sciences* 10, 9 (2020), 3036.

[46] A Guruva Reddy and Srilatha Narava. 2013. Artifact removal from EEG signals. *International Journal of Computer Applications* 77, 13 (2013).

[47] Nigel C Rogasch, Mana Biabani, and Tuomas P Mutanen. 2022. Designing and comparing cleaning pipelines for TMS-EEG data: a theoretical overview and practical example. *Journal of Neuroscience Methods* (2022), 109494.

[48] David Rozado, Andreas Duenser, and Ben Howell. 2015. Improving the performance of an EEG-based motor imagery brain computer interface using task evoked changes in pupil diameter. *PloS one* 10, 3 (2015), e0121262.

[49] David Rozado and Andreas Dunser. 2015. Combining EEG with pupillometry to improve cognitive workload detection. *Computer* 48, 10 (2015), 18–25.

[50] Mohammad R Saeedpour-Parizi, Shirin E Hassan, and John B Shea. 2020. Pupil diameter as a biomarker of effort in goal-directed gait. *Experimental Brain Research* 238, 11 (2020), 2615–2623.

[51] Jonathan Smallwood, Kevin S Brown, Christine Tipper, Barry Giesbrecht, Michael S Franklin, Michael D Mrazek, Jean M Carlson, and Jonathan W Schooler. 2011. Pupillometric evidence for the decoupling of attention from perceptual input during offline thought. *PloS one* 6, 3 (2011), e18298.

[52] Winnie KY So, Savio WH Wong, Joseph N Mak, and Rosa HM Chan. 2017. An evaluation of mental workload with frontal EEG. *PloS one* 12, 4 (2017), e0174949.

[53] Lars St, Svante Wold, et al. 1989. Analysis of variance (ANOVA). *Chemometrics and intelligent laboratory systems* 6, 4 (1989), 259–272.

[54] Petre Stoica, Randolph L Moses, et al. 2005. Spectral analysis of signals. (2005).

[55] M Stone. 1978. Cross-validation: A review. *Statistics: A Journal of Theoretical and Applied Statistics* 9, 1 (1978), 127–139.

[56] Salah Taamneh, Panagiotis Tsiamyrtzis, Malcolm Dcosta, Pradeep Buddharaju, Ashik Khatri, Michael Manser, Thomas Ferris, Robert Wunderlich, and Ioannis Pavlidis. 2017. A multimodal dataset for various forms of distracted driving. *Scientific data* 4, 1 (2017), 1–21.

[57] Yahui Wang, Suihuai Yu, Ning Ma, Jinlei Wang, Zhigang Hu, Zhuo Liu, and Jibo He. 2020. Prediction of product design decision Making: An investigation of eye movements and EEG features. *Advanced Engineering Informatics* 45 (2020), 101095.

[58] D Wildemeersch, N Peeters, V Saldien, M Vercauteren, and G Hans. 2018. Pain assessment by pupil dilation reflex in response to noxious stimulation in anaesthetized adults. *Acta Anaesthesiologica Scandinavica* 62, 8 (2018), 1050–1056.

[59] Zhong Yin and Jianhua Zhang. 2017. Cross-session classification of mental workload levels using EEG and an adaptive deep learning model. *Biomedical Signal Processing and Control* 33 (2017), 30–47.

[60] K Yu, I Prasad, Hasan Mir, N Thakor, and Hasan Al-Nashash. 2015. Cognitive workload modulation through degraded visual stimuli: a single-trial EEG study. *Journal of neural engineering* 12, 4 (2015), 046020.

[61] Pega Zarjam, Julien Epps, and Nigel H. Lovell. 2015. Beyond Subjective Self-Rating: EEG Signal Classification of Cognitive Workload. *IEEE Transactions on Autonomous Mental Development* 7, 4 (2015), 301–310. https://doi.org/10.1109/TAMD.2015.2441960

[62] Minrui Zhao, Hongni Gao, Wei Wang, Jue Qu, and Long Chen. 2020. Study on the identification of irritability emotion based on the percentage change in pupil size. In *Proceedings of the 2020 2nd International Conference on Image, Video and Signal Processing*. 20–24.

[63] Yueying Zhou, Shuo Huang, Ziming Xu, Pengpai Wang, Xia Wu, and Daoqiang Zhang. 2021. Cognitive Workload Recognition Using EEG Signals and Machine Learning: A Review. *IEEE Transactions on Cognitive and Developmental Systems* (2021).