POLISH-JAPANESE ACADEMY OF INFORMATION TECHNOLOGY

Computer Science

Polish-Japanese Academy of Information Technology

Monika Kaczorowska
Student number: 19853

Measurement and Analysis of Cognitive Workload on the Basis of Eye-tracking Activity Using Machine Learning

Dissertation supervisor:
Prof. Adam Wierzbicki, Ph.D.

Dissertation co - supervisor:
Małgorzata Plechawska-Wójcik, Ph.D.

Warsaw, February, 2023

I dedicate this to my husband
for his constant love and support.

Abstract. A rapidly developing world brings up a progressively increasing need for professions which require: high qualifications, strong intellectual capabilities and abilities of coping with intensive cognitive workload. Roles such as those of pilots, professional drivers or air traffic controllers are related to significant cognitive workload and a mistake caused by mental fatigue can cause irreversible damage. Due to this the importance of estimating cognitive workload becomes especially significant. The most valuable data for analysis can be gathered directly in the process of carrying out professional activities. Hence, a practical tool convenient for collecting data for analysis is needed. The most frequently applied method of data collection for cognitive workload estimation is electroencephalography, which is complicated, prone to noise and time consuming. In my research I suggest the application of an eye-tracking technique for data collection combined with explainable machine learning methods for cognitive workload estimation as the solution to these issues.

The aim of my research was to investigate whether features based on eye-tracking and user performance can be used to classify cognitive workload and aid the development of an interpretable machine learning model allowing to classify cognitive workload levels. The improvement of the quality of cognitive workload level classification was also the goal of my study. In order to achieve the goals I have collected the experimental data, developed the processing procedure, and tested it on the collected data. A cognitive workload assessment was performed using machine learning methods using binary and multiclass approaches. All of the machine learning models were developed on the basis of a subject-independent approach. This approach is more general and enables the creation of a more flexible classification model allowing to predict the cognitive workload of any participant. The model is trained on the data of several participants and can be used on another participant.

I have performed a series of cognitive workload studies and I have conducted the following analyses: appliance of interpretable machine learning, appliance of fuzzy aggregation functions and calculation of new features. Interpretable machine learning was used in a multiclass classification task which allowed to analyze the importance of features and to understand the mechanism accompanying the processes related to cognitive load. In the next studies, fuzzy aggregation functions were used, which improved the results of classifying levels of cognitive load. This approach is based on a set of classifiers and the use of aggregation functions made it possible to improve the results in the case of initially weaker results of separate classifiers. An ex-Gaussian distribution was used to calculate new features for a model predicting cognitive load levels. The use of ex-Gaussian modeling was valuable in detecting dissimilarities. In all of the tested approaches, highly accurate classification results of over 96% were achieved. In conclusion, this dissertation may be useful to researchers looking for ideas and techniques for studying cognitive load.

Streszczenie. Szybko rozwijający się świat powoduje coraz większe zapotrzebowanie na zawody wymagające wysokich kwalifikacji, zdolności intelektualnych oraz umiejętności radzenia sobie z intensywnym obciążeniem poznawczym. Stanowiska takie jak pilot, kierowca zawodowy czy kontroler ruchu lotniczego wiążą się ze znacznym obciążeniem poznawczym, a błąd spowodowany zmęczeniem psychicznym może mieć wysoką cenę. W związku z tym znaczenie oszacowania obciążenia poznawczego staje się szczególnie istotne. Najcenniejsze dane do analizy można zebrać bezpośrednio podczas wykonywania pracy zawodowej. Potrzebne jest więc praktyczne narzędzie do wygodnego zbierania danych potrzebnych do analizy. Standardową, najczęściej stosowaną metodą zbierania danych do szacowania obciążenia poznawczego jest elektroencefalografia, która jest skomplikowana, czuła na szumy i czasochłonna. W swoich badaniach jako rozwiązanie proponuję zastosowanie techniki eye-tracking do zbierania danych w połączeniu z wyjaśnialnymi metodami uczenia maszynowego do szacowania obciążenia poznawczego.

Celem moich badań było zbadanie, czy cechy oparte na śledzeniu wzroku i wydajności użytkownika mogą być wykorzystane do klasyfikacji obciążenia poznawczego oraz opracowanie interpretowalnego modelu uczenia maszynowego pozwalającego na klasyfikację poziomów obciążenia poznawczego. Celem moich badań była również poprawa jakości klasyfikacji poziomów obciążenia poznawczego. Aby osiągnąć założone cele, zebrałam dane eksperymentalne, opracowałam procedurę przetwarzania i przetestowałam ją na zebranych danych. Ocena obciążenia poznawczego została przeprowadzona metodami uczenia maszynowego w podejściu binarnym i wieloklasowym. Wszystkie modele uczenia maszynowego zostały opracowane w oparciu o podejście niezależne od badanej osoby. Takie podejście jest bardziej ogólne i pozwala na stworzenie elastycznego modelu klasyfikacji pozwalającego przewidzieć obciążenie poznawcze każdej badanej osoby. Model jest uczony na danych kilku uczestników i może być użyty dla innego uczestnika.

Wykonałam serię badań obciążenia poznawczego oraz przeprowadziłam następujące analizy: zastosowanie interpretowalnego uczenia maszynowego, zastosowanie rozmytych funkcji agregacji oraz obliczenie nowych cech. Interpretowalne uczenie maszynowe zostało wykorzystane w zadaniu klasyfikacji wieloklasowej, pozwoliło na analizę ważności cech oraz zrozumienie procesu towarzyszącego procesom związanym z obciążeniem poznawczym. W kolejnych badaniach wykorzystano rozmyte funkcje agregacji, które poprawiły wyniki klasyfikacji poziomów obciążenia poznawczego. Podejście to opiera się na zbiorze klasyfikatorów, a zastosowanie funkcji agregujących umożliwiło poprawę wyników w przypadku początkowo słabszych wyników poszczególnych klasyfikatorów. Rozkład ex-Gaussa wykorzystano do obliczenia nowych cech modelu przewidującego poziomy obciążenia poznawczego. Wykorzystanie modelowania ex-Gaussa jest cenne w wykrywaniu odmienności. We wszystkich testowanych podejściach osiągnięto wysokie wyniki klasyfikacji, przekraczające 96%. Podsumowując, rozprawa może być przydatna badaczom poszukującym pomysłów i technik badania obciążenia poznawczego.

Table of Contents

1 Introduction

Understanding cognitive workload is of great importance in monitoring human mental fatigue. Cognitive workload is described as a quantitative measure of the amount of mental effort necessary to perform a task [1, 2]. Work is divided into two groups: physical and mental. In case of physical work, it is quite simple to say that someone is exhausted; this is not the case for mental activity. Failure to recognize fatigue caused by excessively high levels of cognitive workload can be extremely dangerous. Mental fatigue has a negative influence on reaction time, and decreases the cognitive system performance of the brain in terms of perception, attention, analyzing and planning. Estimation of cognitive workload is an essential part of professions such as pilots, traffic controllers or drivers, on whom the life and health of many people depends. Mistakes made by such professionals might cause huge material losses as well. What is more, the assessment of cognitive workload capacity might be useful in the process of modeling information processing capabilities. An automatic cognitive workload determination mechanism can help to prevent negative effects of exhaustive mental effort and develop learning techniques.

The presented method is based on artificial intelligence models that enable the classification of cognitive workload levels using eye-tracking signals. Previously, electroencephalography signals were applied in cognitive workload science. Nowadays, eye-tracking signals are used more and more frequently in many fields, not only in detecting cognitive workload levels.

The proposed non-invasive method can be easily adopted. This approach develops subject-independent classification regardless of age or habits of an examined person. Interpretable machine learning was applied, which made it possible to understand the studied phenomenon more deeply, while achieving a high-quality estimation of the cognitive workload level. This increased understanding is of fundamental importance in the development of the state of knowledge about cognitive workload processes, which is of interest to other fields of science dealing with the study of human mental activity.

In contrast to physical effort, which may manifest as pain in the muscles, cognitive fatigue might have various symptoms, varying from person to person. The challenge is to detect the physiological process on the basis of which it will be possible to determine the level of cognitive fatigue. The proposed approach provides a new point of view on the problem of classification of cognitive workload. The development of interpretable machine learning and aggregation functions allows for a more efficient processing of data. Additionally, analysis of the importance of processed features using interpretable machine learning techniques allows for a deeper understanding of mental processes and ensures better performance. This approach enables the analysis of individual components of the model along with the selection of features that best distinguish each level of cognitive workload. Moreover, the application of a group of classifiers, the results of which are calculated with the use of fuzzy aggregation methods such OWA and Choquet, allow to improve the accuracy of the classification results.

In order to carry out the analysis and classification process of cognitive workload, a dedicated experiment was designed and conducted.

1.1 Research Problem

In order to explore the research problem I have formulated the following three research objectives.

1. **Investigating whether features based on eye-tracking and user performance can be used to classify cognitive workload levels.**

   The research goal focuses on an attempt to detect the absence or presence of mental fatigue and obtain the satisfying results using a subject-independent approach. Another aim is to check if the feature selection process allows to obtain better results in the classification of cognitive levels.

2. **Development of an interpretable machine learning model which allows the classification of cognitive workload levels.**

   The research goal is to perform a multiclass subject-independent classification of cognitive workload based on eye-tracking and user performance data using feature selection based on interpretable machine learning models.

3. **Improvement of the quality of cognitive workload level classification.**

   The research goal is to improve the classification model by applying fuzzy aggregation methods. The probabilities of belonging to each class are treated as input to aggregation functions. Ex-Gaussian statistics can be applied in the feature extraction step. Feature analysis is possible through the use of interpretable machine learning models.

2 Overview of the State-of-the-art

2.1 Cognitive workload definitions and application

Understanding cognitive workload gives a great opportunity in understanding human mental fatigue [3]. Cognitive workload is described as the quantitative measure of the amount of mental effort needed to perform tasks [1, 2]. Tasks vary in complexity and require different levels of concentration. In [4, 5] the authors describe mental workload as the relation between the resources required by the task and those available to the human. In [6] mental workload is defined as the cognitive demand of a task. A person can experience cognitive overload when the mental activity performed requires more resources than they have [7]. The assessment of mental effort might be useful in the process ofmodelling information-processing capabilities. The cognitive workload process helps in explaining mental fatigue and its influence on the brain's cognitive system performance in terms of perception, attention and target detection failure [8]. The increase of intensity in work causes mental overload. A high level of cognitive workload requires the participant to use extra resources which can lead to a decrease of efficiency and performance [9]. Edith Galy et. al published their studies [9] where they tested the influence of factors which affect cognitive workload: task difficulty, time pressure and alertness. Both task difficulty and time pressure have an influence on cognitive workload [10]. Another stressful factor in work is time pressure. If several tasks are performed simultaneously and relay on the same source, they interfere with each other and compete for the resource. The brain cannot process all of the information and it could be dangerous in the real world. Each day people may experience several workload levels such as underload, medium and overload but keeping a balanced workload allows people to work safely and effectively [11]. The presented cognitive workload descriptions show that there is not only one proper definition. In general, cognitive workload theory says that people have limited cognitive and attentional capacity and that different tasks require different resources to be processed.

The knowledge of cognitive workload is very important in professions such as those of drivers [12, 13], pilots [14, 15] or traffic controllers [16]. In [12], Patten et al. evaluated the driver experience on cognitive workload in the context of real-driving. The authors were focused on the ability of drivers to manage the additional cognitive workload. Dehais and colleagues monitored pilots' cognitive workload simulating low load when participants were only watching a flight and high load when the participants were piloting an airplane [14]. In [16] the authors presented factors enabling the assessment of cognitive workload in a real traffic controller environment. Cognitive workload recognition is applied in education as well. The assessment of cognitive workload during the learning process helps to understand the complexity of learning [17]. Mental workload studies allow us to understand and improve work conditions in a field of ergonomics [9]. Zhang and colleagues developed a system allowing the development of driving skills in people on the autism spectrum [18]. The assessment of cognitive workload might also be used in a medical context e. g. in [19] the authors conducted an experiment which allowed to estimate the cognitive load among people who were diagnosed with cancer. Ensuring appropriate levels of cognitive workload of surgeons during their work is essential – this fact was shown in the research of Ortega-Moran and colleagues. They examined the cognitive workload of surgeons during surgical interventions and concluded that it should be monitored [20].

Cognitive workload is multidimensional and can be widely applied to assess and understand human resources and to help the quality of work and wellbeing. The terms cognitive workload and mental workload are synonymous in the literature [21].

2.2 Cognitive workload assessment

Mental workload can be measured using different types of tools, which can be divided into three main groups: subjective measures, performance and psychophysiological measures [9]. In order to assess the effectiveness and the difficulty of tasks the following subjective measures can be used: the NASA Task Load Index (NASA-TLX) [22], the Subjective Workload Assessment Technique (SWAT) [23] and the Rating Scale Mental Effort (RSME) [24]. The NASA-TLX scale is an assessment tool which contains six subscales: Mental Demand, Physical Demand, Temporal Demand, Overall Performance, Effort, and Frustration Levels. Each subscale contains the questions which enable to give a correct and true answer. The subscales are rated using a 100 point scale with a 5 point step. In [25, 26, 27] the authors applied the NASA-TLX scale by asking drivers to assess their mental workload. In [13] the authors also applied the NASA-TLX scale to assess the cognitive workload during driving. The SWAT scale contains three aspects of cognitive workload: Time Load, Mental Effort Load and Psychological Stress Load. A participant has to assign 1, 2 or 3 points to each aspect. Zulfany and colleagues [28] applied the SWAT scale to analyze the mental workload of software engineers who work remotely. In [29] the authors applied the SWAT scale in the assessment of cognitive workload in a military environment. The RSME is similar to the NASA-TLX and consists of a line with a 150 points range with a 10 point-step containing nine labels indicating a degree of effort: Absolutely No Effort, Almost No Effort, A Little Effort, Some Effort, Rather Much Effort, Considerable Effort, Great Effort, Very Great Effort, Extreme Effort. A participant is asked to mark how much effort was needed to finish the task. Ghanbary et al. [30] published their research where they applied the RSME scale in the assessment of cognitive workload of nurses. The authors use various questionaries for self-evaluation in their cognitive workload researches, one of them is the Instantaneous Self Assessment (ISA) [31, 32]. In [33] the subjective techniques enabling cognitive workload assessment were compared. Some of the measures were originally designed for air traffic control but nowadays they are widely used in ergonomics, driving or education.

The second way of cognitive workload measurement is to take into consideration the performance of tasks, for instance: responses [34], reaction time, the Inverse Efficiency Score [35] defined as a mean of reaction time by percentage of correct answers. If the requirements for the speed of task execution increase, the accuracy of task execution decreases. Zarjam et al. used responses obtained during arithmetic tasks in their cognitive workload estimation [36]. In [37] the authors used driving performance and psychological measures for cognitive workload analysis. Lobo et al. used performance related features such as correctness of performed tasks [38]. Ktistakis and colleagues measured reaction time and the Inverse Efficiency Score during cognitive workload tasks [39].

Different types of tasks are applied in cognitive workload experiments such as arithmetical tasks [36], silent reading of texts [40] or playing games [41] and simulated driving [42]. In [43] the authors designed a dataset including cognitive load tasks: intelligence test and memory test. The N-back tasks are used to measure working memory capacity and cognitive workload [44, 45, 56]. The N-back approach is based on presenting a sequence of stimuli and the participant is asked to indicate if the current stimulus is the same as the stimulus N steps earlier. The secondary task approach is applied to simulate cognitive workload [39] as well. The participants perform a secondary task while performing a primary task keeping the efficiency of the primary task. In order to induce cognitive workload a modified version of the paced auditory serial attention test [47] (mPASAT) can be applied [48]. It requires a working memory, attention and arithmetic capabilities. In [48] the authors reported that the mPASAT evokes cognitive workload. The Multi-Attribute Task Battery (MATB) [49] is employed in cognitive workload studies. The MATB includes analogous tasks to activities which a pilot performs during a flight [50]. In [51] the authors proposed the Oculo-Cognitive Addition Test (OCAT), which tracks eye movements when users carry out mental addition tests in three levels of cognitive workload: low, medium and high.

The Digital Symbol Substitution Test (DSST) is a cognitive tool which allows to understand human associative learning [52]. Originally this tool was created in a paper-and-pencil version. It is a well-known tool which enables to check and verify memory, a patient's processing speed and the executive functioning of the examined people. It is widely applied in clinical neuropsychology to measure cognitive dysfunction [53, 54].

Physiological measures are widely applied in mental effort estimation enabling contiguous monitoring and measurement of cognitive workload. However, physiological measures are sensitive to physical effort and it is recommended to use them in case of a small physical effort [55].

The literature review has shown that cognitive workload can be measured based on bio-signals such as brain-activity [56, 57], eye activity [25], pupillometry [58], functional galvanic skin response (GSR) [59], heart activity and blood pressure [60, 61]. Brain activity can be measured using electroencephalography (EEG) [62] or functional near-infrared spectroscopy (fNIRS) [63]. In the EEG technique the brain activity is recognized when the difference between electrode and neural signal appears while in the fNIRS method, it is recognized by using near-infrared light to measure the concentration change of oxygenated and de-oxygenated hemoglobin which is immune to electrical noise. Electroencephalography has been proven to be useful in cognitive-related sciences:[14, 40, 44, 46]. In [64] the authors reviewed the EEG-based cognitive workload recognition using machine learning.

Eye-activity is a non-invasive method which may be measured by using an electrooculograph, an eye-tracker or a pupilometer [65, 66]. The EOG technique reflects eye-movement using the electrodes commonly placed above and below the eye. When the eye moves, the potential difference appears and the signal called electrooculogram can be observed [65]. An eye-tracker allows to monitor eye-movements while a pupillometer enables to measure the size and reactivity of the pupil (pupillometry) which are considered to be the indicators of mental workload as well [37, 66, 67]. In [68] the authors examined the cognitive workload of firefighters during searching and rescuing. The metric of task performance was the number of victims found by the crew and the time needed to rescue them. Eye-tracking data were analyzed to find gaze patterns and assess cognitive workload. In [67] the authors conducted studies by using the measurement of pupil oscillation data to estimate cognitive workload levels. Duchowski et al. [67] conducted studies where participants had to solve arithmetic tasks of three different levels. The authors measured the working memory capacity for each participant using the Digit SPAN task (DSPAN) [69]. The DSPAN task is based on recalling the digits in the order in which they appeared (forward-span) or in the reverse order (backward-span).

The task is considered to be completed when the participant makes mistakes in two sequent trials. The authors asked the participants to fill in two questionaries: the NASA-TLX and the Self-Assessment Manikin (SAM), which enabled to assess their emotional valence [70]. The authors [67] proposed a metric called the Index of Pupillary Activity (IPA) and wrote that the IPA can discriminate between task difficulty in the context of cognitive workload. They suggest that the IPA could be sensitive to task difficulty independently of the working memory capacity.

The GSR measurements relay on detecting changes in electrical activity coming from changes in the sweat gland. The GSR technique uses electrodes which must be sensitive to these changes [71]. The sweat gland activity reflects the intensity of the emotional state of the examined person. Nourbakhsh et al. [72] examined temporal and spectral features of the GSR againsttasks of varying difficulty. On the other hand the authors in [73] used the GSR as a cheap tool for measuring mental workload in their research but claimed that it is non sensitive in distinguishing different levels of cognitive workload. Cardiac activity measures are applied in the assessment of cognitive workload because they are relatively inexpensive, do not require advanced training to collect data and they are resistant to the participants' movement [74]. Cardiac activity measures include, among others, heart rate, heart period which is defined as the average time in milliseconds between heartbeats, and blood pressure [74]. The researchers sometimes apply more than one type of bio-signals to measure cognitive workload [75, 76, 77]. Additionally, they use subjective measures in combination with task performance or bio-signals [78, 79]. In [80] the authors used the fNIRS signals and eye-tracking signals to examine cognitive workload during driving.

2.3 Eye-tracking

Eye-tracking is a powerful tool in measuring the point of gaze and eye motion [81]. Charles Bell is the pioneer of eye-tracking research who first described that a physiological connection between the eyes and the nervous system connected eye movement with neurological and cognitive processes [82]. An eye-tracker allows to measure where and in what order a participant is looking during a specific task. Cognitive processes such as perception, memory, language and decision making stimulate where a person looks [83]. Eye-tracking enables to observe the whole cognition in real time and not only the final result of cognition. People usually cannot control eye movements constantly and are not good at remembering where they looked [84]. Due to that eye-tracking is widely applied in studies dedicated to mental processes. The use of eye-tracking has been widespread within the last 20 years and it found its applications in psychology, medicine, neuroscience, computer science, education [85], linguistics and other areas [83]. The literature overview shows examples of the use of eye-tracking in various fields of life such as medicine [86, 87, 88], spelling disorders [89, 90], user experience design [91], consumer research [92], aviation [93], transportation [94], navigation [95], education and software engineering studies [96, 97]. One of the research areas where eye-trackers are often used is related to car driving [98]. In [89] the authors used eye movements to examine the disorder of spelling awareness. In [99] the authors performed a literature overview covering the applications of eye-tracking studies in aviation, maritime and construction. The eye-activity data turn out to be useful in cognitive workload analysis [20, 65, 100].

An eye-tracker can be used by disabled people to control a wheelchair [101] as well. Eye-tracking data are useful in analyzing cognitive workload [102] allowing to detect neuropsychological diseases or mental fatigue [103, 104, 105]. Moon et al. examined cognitive workload of people with Parkinson’s disease using eye-activity [106].

In [107] the authors conducted a literature review about surgery where eye-tracking was used to assess the cognitive workload of the participants. The studies showed that eye movements such as pupil responses, gaze patterns and blinks are indicators of cognitive workload from the viewpoint of surgery. Gil and Amalia [108] presented a literature review about application of an eye-tracker in surgery [108].

Nowadays, eye-trackers usually have the form of video-based tools identifying pupil parameters. Eye-trackers usually use near-infrared light which is invisible to people. An eye-tracker sends near-infrared light which is reflected from the eye and then captured by the cameras on the device. The data are processed using dedicated algorithms that capture the details of a participant's eyesight and calculate the focus point on the screen. A calibration procedure has to be carried out before the measurement of the eye-position. It is a process which allows to estimate the characteristic of a participant's eyes and optimize the eye tracking algorithm [83]. In [109] the authors showed that a webcam could be used as a low-cost alternative to an eye-tracker. Eye-tracking data in raw form are a series of samples. Raw eye-tracker data are rarely processed by themselves. More often the measures describing eye movement are extracted for further analysis [97].

The eye-tracking feature set consists of fixation related features, saccade related features, blink related features, and pupil related features [83]. According to [110] fixations are defined as the period of uptake of visual information when a participant's eyes are in a stable position. Single fixation appears between two sequent saccades. Fixations vary in length depending on visual stimuli, difficulty of the task, skills and attention. They usually last 180-330 milliseconds [110]. The eye cannot obtain visual information from single fixations so the eyes have to move frequently and the fixations are relatively short [111]. Even during fixations, eye movement can cause drift and microsaccades [112]. In [110] saccades are described as a rapid movement between two sequent fixations. The gaze moves from one point to another bringing the part of visual information. Similarly to fixations, the size and duration of saccades can be different depending on the tasks being performed [111]. Typically they last 30-50 milliseconds. Other types of eye-movement are smooth pursuit or vergence which can be made deliberately like saccades [83].

An eye-tracker enables to register changes in diameter of the pupil as well. Pupillometry is the measurement of pupil size which can change as the response to illuminance and cognitive processes [113, 114]. Blinks also appear in the eye-tracking process [115] and they are identified as zero data occurring in two saccadic events [116]. In [117] the authors presented eye-tracking indicators and their role in cognitive workload measurement. Joseph and Murugesh published a very clarified description of eye-tracking related metrics, such as saccades, pupil dilation, and scan path and their applications in cognitive workload estimation [117]. In [118] the authors compared microsaccade metrics to pupillometric measures. The authors reported that the microsaccade magnitude could be related to the task difficulty in cognitive workload.

Features related to fixations and saccades are commonly used in eye-tracking analysis applied in psychology and neuroscience [119]. In [104, 120, 121], the authors use these features to analyze mental fatigue, behavior patterns and disorders such as schizophrenia and autism.

The following features related to fixations and saccades can be presented: total number of saccades, mean duration of saccades, total number of blinks and mean duration of a blink. Kardan at al. used standard deviation, mean and skewness of fixation duration as well as saccade amplitude in their classification of mental states [122]. Standard deviation of saccades was calculated by Dowiasch and colleagues in eye-movement analysis depending on the participants' age [123]. In [98] the authors examined a method based on fixation-aligned pupillary response averaging in a driving context. Li and colleagues [124] carried out a literature review about eye-tracking parameters and methods allowing the assessment of mental states. In [125] the authors presented a summary of eye-tracking related features applied in emotional and cognitive process detection. Guo et al. [126] wrote that fixation duration, saccade frequency, saccade duration, pupil diameter and pupillary activity can be considered as features to estimate cognitive workload. Blink related features have also been proven to be an indicator of cognitive processes [25]. Pupil related features such as pupil dilation are considered as cognitive workload measure as well [78].

2.4 Ex-Gaussian statistics in eye-tracking data

Ex-Gaussian statistics are used in order to model the process of reaction to a stimulus. The ex-Gaussian distribution allows to calculate three parameters: μ – corresponding to the mean of the normal component, σ – corresponding to the symmetric standard deviation of the normal component and τ – corresponding to the exponential part of the distribution [127]. The Ex-Gaussian method is a common in reaction-time studies. Otero-Millian et al. conducted one of the first studies using the ex-Gaussian method for eye-tracking data [128]. The distribution of variables such as fixation length and intersaccadic intervals were close to the ex-Gaussian distribution. A literature overview about the application of distributional analyses to fixation durations based on the ex-Gaussian method was presented in [129]. The authors [129] have shown the advantages of ex-Gaussian modeling in fixation duration modeling. It can be said that the ex-Gaussian distribution is a proper model for fixation distribution. Guy at al. [129] presented the parametric mean of fixation duration and published evidence that show that empirical distribution of fixation duration is sensitive to task input. The authors [129] reported that the parameter τ is associated with repetitive exposures to the same images but parameter μ is associated with stimuli familiarity and efficiency of solving tasks. Components of the ex-Gaussian distribution can reflect cognitive processes. The increase of parameter τ might be related to attentional lapses regarding repetitive presentation of a stimulus. Luke and colleagues [130] applied the ex-Gaussian modelling to examine the differences between fixation duration distributions and working memory capacity. Ex-Gaussian statistics seems to be a good tool for dissimilarity detection in groups [131].

2.5 Cognitive workload classification

Cognitive workload is a subject of numerous studies especially in the aspect of classification of cognitive workload levels [38, 64, 132]. A literature overview shows that the majority of cognitive workload studies concern binary classification problems [104, 133, 134], and that multiclass classification studies are less common. The presented models distinguish between low and high levels of cognitive workload [135]. In addition to binary classification, a multiclass classification can be found, e.g. a three-class classification distinguishing between low, medium and high levels of cognitive workload [38, 136, 137].

Researchers conduct their classification process in two ways: subject-dependent [133, 132] and subject-independent approaches [136, 135]. In the subject-dependent approach the classification model is trained on the data of a given participant. The results of [132] show that the subject-dependent model allows to obtain higher classification performance than the subject-independent approach. However, a subject-independent approach is more general and enables to create a more flexible classification model allowing to predict the workload level of any participant. The development of the subject-independent approach called cross-subject allows to classify the levels of cognitive workload regardless of age or habits of the examined person [136]. The model is trained on the data of several participants and can be used on another participant. The subject-independent approach seems to be more attractive nowadays [104] because this approach allows to create a general model to estimate cognitive workload. Examples of combining the subject-dependent and the subject-independent approaches can be found in the literature as well [132, 138, 139]. In [140] the authors claim that the benefit of using the cross-participant approach seems to outweigh the better efficiency in the subject-dependent approach because it requires less effort in calibration and training.

Researchers apply both classical models such as the Support Vector Machine (SVM) [104, 34], Linear Discriminant Analysis (LDA) [141], k – Nearest Neighbors (KNN) [38, 141, 142], Random Forest [34, 137], Multilayer Perceptron (MLP) [143], Regression models [139], Naive Bayes [136] and deep neural networks such as convolutional deep neural networks [134, 144]. Researchers often report accuracy [143, 145] in their studies.

In terms of classification of cognitive workload levels, the training data have to be labeled as in supervised learning. The most known approach is to define the difficulty of tasks by an expert [133, 19]. Another approach is to use a Rash model [146] or a stress-strain model [147, 148] which allow to adjust the cognitive workload level to each participant separately [137]. Baldwin at. al. [149] mention that two people might be able to perform tasks but with a different workload. However, most of cognitive workload research studies apply an approach based on difficulty split by an expert [36, 39, 141, 150].

Classification of cognitive workload can be carried out based on psychophysiological signals such as electroencephalographic data (EEG) [64, 136], eye-tracking data [104], galvanic skin response (GSR) [59] or heart rate [141]. The studies of classification of cognitive workload with the combination of two bio-signals such as EEG and eye-tracker can be found in the literature as well [135]. In [141] the EEG, EOG and ECG signals were used in the classification of drowsiness levels. Wobrock and colleagues [57] tested a model predicting mental effort using the EEG, ECG and GSR methods.

Cognitive workload classification studies have two purposes. The first goal is to find an indicator of cognitive workload, for example, cognitive or psychological while the second goal is to verify and develop computational methods allowing to predict the levels of cognitive workload.

2.6 Cognitive workload classification based on non-eye-activity data

The literature overview shows that the estimation of cognitive workload levels can be performed based on non-activity data [59, 133, 136, 141] especially based on EEG signals. In [136] the authors presented a subject-independent approach in cognitive workload classification based on EEG signals using hierarchical Bayes. The three-level classification of cognitive workload: low, medium and high was performed. Eight participants took part in the experiment. They were asked to perform the Multi-Attribute Task Battery. A feature selection process was done using discrete-time short-term Fourier transform (STFT). The authors reported mean classification accuracies for separate people from 0.42 to 0.8. Khushaba et al. [141] examined 31 people to study the drowsiness of drivers. They asked them to perform a driving simulation task. Based on features extracted from the EEG, EOG and ECG signals, they created the classification model which is able to estimate one of five drowsiness levels. The authors mentioned that EEG is an essential element of their analysis because EEG helps in detecting drowsiness. The author tested four different classifiers: LDA, kNN and SVM in two implementations (LIBLINEAR and LIBSVM). The researchers published that they obtained 95%-97% of accuracy across all subjects. Based on EEG signals the authors in [143] created a multiclass classification model to predict one of seven cognitive workload levels using the Multilayer Perceptron. 12 participants took part in the experiment. Their task was to solve arithmetical tasks [36]. The authors tested the model by using the leave-one-out technique, across all participants and obtained 97.53%-98.62% of accuracy. The authors used extracted features based on wavelet-based features. Supler and colleagues [139] published another scientific work, presenting the application of EEG signals for prediction of cognitive workload levels. They asked 10 participants to perform arithmetical tasks with increasing levels of difficulty using Q-value [151] which is used to assess the difficulty of arithmetic tasks. They tested the subject-independent and the subject-dependent approaches using the Linear Regression Model. The classification accuracy for three defined difficulty levels: easy, medium and difficult based on all datasets was accordingly 44.8%, 74.2% and 30.4%. These results were reported for classification across all of the participants. The authors tested different electrode subsets obtaining the average accuracy between 45.5% and 55.8%. The authors conducted within-participant estimation using 10-fold cross-validation and reported the average correlation coefficient (CC) 0.9, root mean squared error (RMSE) 0.95 while the cross-participant prediction CC was equaled to 0.82 and the RMSE to 1.34. Hefrom et al. applied convolutional recurrent neural networks to predict cognitive workload: low and high. They tested four approaches based on the subject-independent approach using six models such as Long Short-Term Memory (LSTM), multi-Path Convolutional Recurrent Neural Network (MPCRNN) etc. The models were trained based on EEG data gathered from eight participants during MATB tasks. The author reported the best accuracies for MPCRNN from 79% to 86.8% [134]. In [152] Hajinoroozi et al. proposed a channel-wise neural network (CCNN) dedicated to EEG signals to predict the cognitive performance of drivers: good or poor. They tested their algorithm based on data gathered from 37 participants performing cognitive tasks related to driving in a simulator. The authors tested their proposed algorithm with others classifiers such as SVM, LDA, Convolution Neural Network (CNN) on raw EEG data, time-frequency power spectrum and Independent Component Analysis (ICA)-transformed data. Within-subject and cross-subject approaches were tested. The authors reported 86.08% of accuracy for subject-dependent and 63.39% for subject independent approaches for CCNN, trained based on raw EEG data. Wobrock et al. [59] created a model which allows to estimate cognitive workload during 3D modelling. They wrote that applying combined physiological signals: EEG, EMG, ECG and GSR may cause poor performance, most probably due to the curse of dimensionality. They applied the subject-dependent approach and obtained an average accuracy of 88.6% for the combination of EEG and EMG in a binary classification [59]. Almogbel et al. published two scientific works [133, 153] related to detection of cognitive workload of drivers based on EEG signals. The authors tested binary and multiclass classifications based on recordings of 24 hours from 1 participant. The participant was asked to play a computer racing game. The convolutional neural networks were applied and enabled to obtain a promising accuracy from 93.4% to 97.6%. In [154] the authors published a model based on the cross-subject approach which allows to detect the difficulty of arithmetic tasks. The publicly available dataset MIT PhysioNet containing recordings from 36 participants while performing arithmetical tasks was used in studies. The difficulty level: easy or difficult was considered on the basis of correct answers per minute. The feature set consists of entropy, energy and mean of all the sub-bands. The three classifiers : SVM, Decision Tree and Quadratic Discriminant (QD) were tested, achieving the best accuracy of 92% for QD. Chakladar et al. [155] also conducted a two-class classification based on EEG data from publicly available datasets. The model predicts two states: arithmetical tasks or rest state achieving 87% of accuracy. The authors used Long Short-Term Memory (LSTM) and Filter Bank Common Spatial Pattern (FBCSP). In [156] the authors created models based on the subject-independent approach to predict one of two cognitive workload levels: low or high. Becerra-Sánchez and colleagues [156] collected EEG data from eight participants who were using a driving simulator. They asked the participants to fill in the NASA scale TLX and the Instantaneous Self-Assessment (ISA). In [156] the authors proposed a new feature selection model for pattern recognition based on information from EEG – Genetic Algorithms and Logistic Regression for Structuring of Information (GALoRIS). The authors reported over 96% of accuracy using SVM. Han et al. [157] were focused on detecting one of the mental states of pilots: distraction, workload, fatigue and normal state. They took into consideration not only EEG data but also ECG, respiration and electrodermal activities gathered from eight participants. The authors tested several classifiers such as kNN, SVM, LDA, LSTM as a benchmark and proposed a method based on multimodal deep learning with convolutional neural network and long short-term memory achieving 85.2% of accuracy.

The following other scientific works [158, 159, 160] have been published recently and are related to cognitive workload classification based on EEG data. In [158] the authors conducted binary classification of cognitive workload obtaining the best results for Deep Recurrent Neural Network (RNN) achieving 92.8% of accuracy. Taori et al. [159] reported up to 97.8% of accuracy in multiclass classification using the hidden Markov model (HMM). In [160] the authors applied neural networks proposing a multilayer autoencoders ensemble to estimate cognitive workload.

2.7 Cognitive workload classification based on eye-activity data

In recent years, the analyses of cognitive workload based on eye-activity data have become progressively more popular [117, 135, 161]. In [104] the authors published studies of classification of cognitive workload based on eye-tracking data. The eye-tracking data were gathered from 20 participants: young and older adults. The authors asked participants to fill in questionaries asking questions concerning the feeling about mental and physical fatigue using a numerical rating scale. Yamanda and collegous [104] argued that changes in eye-tracking processes in older adults have been reported by other researchers. The participants were asked to perform cognitive tasks based on a modified version of the paced auditory serial attention test [162] (mPASAT). Before and after performing cognitive tasks, the participants watched some video clips. The extracted features consisted of saccade related features, fixation related features, blink related features, pupil diameter related feature, time-series of gaze allocation related features, eye-movement direction related features and saliency-based features [38], in total 181 features. The authors created a classification model which predicts one of two levels of cognitive workload: fatigue and non-fatigue states. The various feature sets were tested and a feature extraction was applied using the Support Vector Machine Recursive Feature Elimination (SVM-RFE) [145]. In order to distinguish between fatigue and non-fatigue states the Support Vector Machine classifier was applied. The best model based on the subject-independent approach achieved 91% of accuracy. Appel and colleagues [163] discussed the subject-dependent approach and mentioned that classifiers are usually too customized to participants based on which were trained. The authors applied the group of the trained classifiers which allow to classify cognitive workload of new participants. These classifiers are used in a weighted voting system. The researchers used a dataset of stimulus based on n-back tasks published in [164]. The authors attempted two and three class classification of cognitive workload levels applying cross and within subject approaches based on pupil data gathered from 25 participants. The following features were extracted: median and maximum of pupil diameter, average blink duration, number of blinks per minute, and the index of Cognitive Activity-events per minute. The index of Cognitive Activity (ICA) was described in [165] as the measure for detecting changes in pupil diameter. In classification studies Extra-Trees were applied because of lower overfit tendency and less computational complexity. Additionally, this model enables to obtain feature importance. The authors wrote that the highest feature weight had median pupil diameter and the least feature weight was assigned to average blink duration. The low and high cognitive workload levels are the most distinguishable in subject dependent and independent approaches. The results of classification of two levels of cognitive workload are 69.8% - 82.4% of accuracy for the subject-dependent approach and 54%-76.8% of accuracy for the subject-independent approach.

In [161] the authors classified cognitive workload of drivers under critical situations. They created a model which predicts one of two cognitive levels: low and high, based on the cross-participant approach. 16 participants took part in an experiment, where the participants were asked to drive in a simulator based on virtual reality (VR) and augmented reality (AG) technology reflecting real driving. The authors labelled the cognitive workload levels: low or high based on defined critical time frames. The set of extracted features included pupil diameters and performance measures such as inputs of accelerator pedal, brake pedal and steering wheel. A two class classification was tested with various classifiers: SVM, Decision Tree, Random Forest and kNN. Each participant was evaluated on a trained model based on the rest of the participants. The best classification result was obtained by SVM – 80.7% of accuracy. The authors reported not only accuracy but also precision, recall and F1-score. Fridman et al [166] carried out an extensive experiment including 92 participants. Data in the form of face images were collected during an on-road experiment based on N-back tasks on a highway. Two models in the subject-independent approach were trained: Convolutional Neural Network 3D (3D-CNN) based on an image and the Hidden Markov Model (HMM) based on pupil position and blink state extracted from an image. In the three class classification the model based on 3D-CNN obtained better results, the accuracy reached 86.1% [166]. In [142] the authors performed three class classification based on 9 features extracted from an eye-tracker. The feature set includes features such as pupil diameter change, number of saccades, saccade duration, number of fixations, fixation duration, number of blinks, blink duration, 2D entropy, 3D gaze entropy. They created a model based on the cross-participants approach which enables to predict one of three cognitive levels: low, medium and hard level of cognitive workload during driving in a simulator. 36 participants took part in the experiment. The participants were asked to perform three scenarios prepared by The Society of Automotive Engineers and three multi-modality secondary tasks. The NASA-TLX scale was used in the experiment to show the influence on mental workload and based on this scale after each scenario the data were labelled with one of three labels: low, medium or high. Chen at el. applied the kNN classifier to create a model achieving 88.9% of accuracy based on 5 features: pupil diameter change, number of saccades, saccade duration, number of fixations and 3D gaze entropy. Farha and colleagues [167] published a study about using eye-tracking data in the assessment of cognitive vigilance. The model predicts one of two levels: alertness and vigilance during Stroop Color Word Task (SCWT). Various classifiers were tested and the higher accuracy of 76.8% was reported for SVM. The author proposed a preprocessing pipeline to process the eye-tracking data: baseline and artifacts correction. The set of extracted features consists of fixation duration, pupil size, saccade duration, saccade amplitude, saccade velocity and blink duration. In [168] the authors presented the method of estimation of cognitive workload during surgical tasks. Wu et al. [168] published a research paper claiming that eye-tracking metrics are sensitive to cognitive workload changes. The authors carried out an experiment asking eight trainees to perform simulated tasks. Eye-tracking features, performance scores and the NASA-TLX rate were taken into consideration in this study.

The model based on the Naive Bayes algorithm predicts low or high level of cognitive workload based on 9 features: sex, trainee level and seven eye-tracking features consisting of pupil diameter mean and standard deviation, gaze entropy, fixation duration and eyelid closure percentage. The authors noticed that pupil diameter and gaze entropy allow to differentiate cognitive workload levels. If the level of cognitive workload increases, these two metrics increase as well. The authors reported 84.7% of accuracy in the subject–independent approach. Bitkina and colleagues [169] presented a scientific paper about the classification of cognitive workload during driving based on eye-tracking metrics. The authors extracted the following eye-tracking metrics: gaze fixation, duration, pointing, and pupil diameter and reported that gaze pointing, fixation duration and pupil diameter are good indicators of driving workload. The Logistic Regression Model was applied to predict low or high workload levels across 7 NASA-TLX categories achieving 83.5% - 95.9% of accuracy. Each participant was asked to fill in the NASA-TLX questionnaire and then label the data: low or high levels were based on the NASA-TLX scores. Zahabi et al. published a scientific work dedicated to the classification of mental states during driving based on the drivers' behavior and eye-tracking measures [170]. In [34] the researchers published their study concerning the evaluation of interfaces for applications in the context of the users' cognitive workload. They carried out an experiment engaging 50 participants between the ages of 20 to 60+ to perform five different tasks on mobile phones. The authors conducted a binary classification predicting low or high levels of workload and a multiclass classification consisting of 3 classes: low, medium and high and 5 class classification: very-low, low, medium, high and very-high. Cognitive workload levels were measured based on behavioral measures such as time and number of steps taken on a task, tasks completed and eye-tracking related features such as number of fixations and saccades, fixation and saccade rate, average, standard deviation and maximum fixation duration, average and standard deviation of pupil dilatation. Among the eight classifiers, Random Forest, Support Vector Machine and k-Nearest Neighbors were tested for each participant. The SVM-Recursive feature Estimation method was applied in the feature selection process. The best results for each type of classification were achieved by Random Forest: 86.8% for 2 class, 74% for 3-class and 62.8% for 5-class. The authors noticed that ageing has an influence on cognitive workload during task performance on mobile phones. The following features are considered as valuable parameters for cognitive workload: age, fixation and saccade number, average of pupil dilatation. The performance measures included Reaction Time and the Inverse Efficiency Score. Ktistakis et al. [39] pointed out that there are few publicly available datasets containing cognitive workload observations. They carried out an experiment involving 47 participants who performed tasks with different levels of difficulty and of a varying duration related to four cognitive workload levels. The participants were asked to find a selected object across nine puzzles and as the secondary task they were asked to perform arithmetical operations. The authors created a dataset called COLET – Cognitive Workload Estimation based on eye-tracking data. The eye-tracking feature set includes fixation related features, saccade related features, peak saccade velocity features, blink related features, pupil diameter features. The NASA-TLX scale was used to assess the cognitive workload level. The following classifiers were tested: Random Forest, Linear Support Vector Machine, Ensemble Gradient Boosting, Logistic Regression etc. The authors labelled data based on two approaches: each activity was related to one cognitive workload level, in total 4 levels, the labels were assigned by using scores from the NASA scales, in total 3 levels: low, medium and high. The authors tested binary and multiclass classifications: 3-class and 4-class of cognitive workload level for two mentioned approaches. The binary classification was performed for each pair of classes e.g, low vs high, low vs medium, first task vs second task etc. The achieved results were among 52% to 98% of accuracy based on cross participant approach getting the accuracy over 88% for 2 class classification: low and high level of cognitive workload. In [171] the authors examined cognitive workload of operators in an oil refinery using eye-tracking data. Shi et al reported that they were able to predict the cognitive workload using the Logistic Regression model.

2.8 Cognitive workload classification based on combination of eye-activity and non-eye-activity data

In the literature cognitive workload classification studies presenting the combination of eye-activity data and other types of data can be found [135, 152]. Lobo et al. [135] classified cognitive-workload based on eye-tracking and EEG data. They performed multiclass classification based on the subject-independent approach predicting one of three levels of cognitive workload. They carried out an experiment where 21 participants got involved and were asked to perform tasks. 21 features were extracted: alpha related features, theta related features, eye-related features such as right eye closure, left eye closure, right pupil diameter, left pupil diameter, performance related feature – correctness [38]. One of the tested approaches was that one participant was used in the test while others were used in the training process. The authors applied kNN and obtained a global average F-score equaling 0.332. The authors attempted to transform the multiclass classification into a binary problem joining low and medium levels together or medium and high together. In [152] the authors conducted an experiment involving 14 participants: both experts and novices in operating a military land platform in normal and dangerous conditions. They wanted to investigate the behavioral and neurophysiological differences among novices and experts. The data were gathered by using fNIRs and a mobile eye-tracker. The eye-tracking features include fixation durations and saccadic amplitudes while based on fNIRs data the following features were extracted: oxygenated hemoglobin (HbO), deoxygenated hemoglobin (HbR) and total hemoglobin (HbT). The combination of HbT and fixation duration features allowed to obtain 91% of accuracy in the two class classification using LDA based on the subject-independent approach. Ziegler et al. [41] published studies based on the subject-independent approach to estimate cognitive workload. 35 participants took part in the experiment in which they were asked to play two games with different levels of difficulty. The authors measured EEG, Heart Rate, and eye movements and pupillometry using an eye-tracker. The distribution of scores obtained in the games played on the easy and hard levels were compared. The SVM and Deep Belief Network were used in the research. The authors proposed to create a group of classification models where each model was trained on similar data. It can ensure high accuracy without creating each model for each individual participant. They doubt that one general classification model can properly estimate cognitive workload. In [172] the researchers proposed a similar approach as in [163], combining the subject-independent and the subject-dependent approaches. The subject-specific models were used as general classifier across participants for new participant data. Each subject-specific classifier predicts the cognitive workload level and then similarities between participants are calculated and classifier weights are taken into account. The models are trained based on physiological data: eye tracking data, heart rate and behavior data. The data were collected from 47 participants during a real-time emergency simulation game containing scenarios of three difficulty levels where participants had to save people from emergency situations. Apart from eye-tracking related features such as fixations, saccades, pupils, heart rate and the game related features defined as actions per second were extracted. In total eight features were used. Two class classification based on Extra-Trees was performed allowing to predict low or high levels of cognitive workload. The mean accuracy for the subject-independent approach equals 70.14%-79.03%, while the mean accuracy of the cross-participant classification is between 67.92% - 80.56%. Cross-validation was applied in the evaluation process. In [150] He and colleagues estimated the cognitive workload of drivers using eye-tracking, heart rate and galvanic skin response. Six eye-tracking features were eye-closure raw data, blink duration, blink frequency, pupil diameter, eyeball rotation speed and PERCLOS – the percentage of eyelid closure over pupil. The authors examined machine learning models in estimating one of three cognitive workload levels. 33 participants took part in experiment and they were asked to perform tasks of various difficulty levels: no difficulty, lower difficulty and higher difficulty based on the n-back task approach. The well-known classifiers such as kNN, SVM, Random Forest, Feedforward neural Network were tested in the subject-independent approach. The author reported that the the most accurate results were obtained based on the combination of eye-tracking, heart rate and galvanic response skin features achieving 97.8% of accuracy for Random Forest. Classification of cognitive workload in driving is one of the most frequently discussed topic recently in [173] as well.

2.9 Interpretable machine learning classification

Nowadays, machine learning is widely applied in various fields [174] such as object recognition, verification and detection [175, 176], classification, analysis or recommendation. One of the methods based on machine learning is applying interpretable machine learning. Interpretable machine learning methods allow to perform feature analysis and understand the classification process. In [177] the authors predicted mental workload using an interpretable machine learning model based on the Gaussian Process Regression and the Multiple Linear Regression. EEG signals were gathered during the experiment. The authors stated that interpretability allows to carry out feature selection which leads to notable decrease of computational complexity. Cui et al. [178] presented a scientific paper related to the convolutional neural network in the detection of drowsiness in drivers based on EEG signals. They applied the interpretable Convolutional Neural Network to find similar EEG features across all participants. The authors used the Class Activation Map [47] to detect the regions in the signal which have the highest impact on classification and thanks to this technique the feature can be analyzed. The authors reported 73.22% of accuracy in the subject-independent binary classification. In [179] the authors applied an interpretable machine learning model based on the Deep Convolutional Neural Network using EEG signals to analyze and understand which cortical regions are more relevant in hand movements. In [180] the authors used interpretable machine learning to classify brain states during visual perception using deep learning models. The authors applied the Shapley Additive Explanations [181] technique for feature importance estimation allowing to create feature rankings. This method gives a great explanation of the classification results for deep learning models. Moreover, the use of interpretable machine learning does not interfere with the use of other approaches in machine learning. The fuzzy aggregation operators can be applied to improve the quality of classifiers taking part in classification process.

Such approach allows to take into consideration results of several classifiers returned separately and then the use of proper aggregation functions enables to aggregate their results [182]. As an aggregation function, one can use various types of mean, Choquet integral [183, 184] or triangular norms [185]. The aggregation functions are applied in many problems for example in face recognition problems [186, 187].

2.10 Conclusion from the State-of-the-Art Overview

The literature review shows an extensive interest of the scientific community towards cognitive workload studies. The topic of cognitive workload estimation is important nowadays and can find its application in many areas. Researchers pay significant attention to various aspects describing cognitive fatigue. As it was observed in the course of the literature review, features for cognitive workload estimation are extracted mainly from EEG signals. Electroencephalography together with its advantages has several features which make it inconvenient in practical use: the signals registered by EEG are highly prone to noise pollution. Moreover, its practical application is highly complicated due to the fact that the preparation for signal registration can last over 30 minutes. Owing to that fact, I decided to use an eye-tracker as the primary source of features for analysis.

Various approaches to cognitive workload estimation can be found, but the common part of the vast majority of the reviewed articles is the following: authors tend to emphasize the quantitative results of cognitive load estimation and prediction. Often the main contribution of an article consists of achieving high values of a chosen machine learning metric. This can be especially notable in case of newly emerging deep learning models that allow to achieve high quality results, but work as black-boxes, rendering their understanding highly complicated. In contrast with the reviewed approaches, the present work extensively applies possibilities of interpretable machine learning models for deeper understanding of the cognitive processes underlying the phenomenon of cognitive workload. Thanks to interpretable models, I was able to indicate the most valuable features that have the highest impact on classification. Another valuable contribution of my work is the fact that I have utilized the subject-independent approach which, combined with interpretable machine learning, can help in understanding the most general patterns accompanying cognitive fatigue. To the best of my knowledge, there are no other examples of application of ex-Gaussian statistics for describing eye-tracking features applied for cognitive-workload estimation. I have decided to use ex-Gaussian statistics to analyze not only temporal eye-movement features but also non-temporal features, such as saccade amplitude. Fuzzy aggregation is a tool which allows to affectively combine the predictions of various classification models. To the best of my knowledge, my work presents the first example of fuzzy aggregation in cognitive workload estimation.

3 Contributions

The contributions concerning Informatics and Cognitive Science are presented in section 3.1 and 3.2. All contributions to Informatics in section 3.1 are related to the classification process of cognitive workload levels based on eye-tracking data. Contributions to Cognitive Science described in section 3.2. are related to eye-tracking feature analysis. The scientific papers describing these contributions in detail can be found in section 6, where they are attached in their full versions.

3.1 Contributions to Informatics

1. **Method for subject – independent classification of cognitive workload based on eye-tracking and user performance features**

   This contribution refers to objective 1. My research has shown that eye-tracking and user performance features based on Digit Symbol Substitution Test (DSST) can be used to correctly classify cognitive workload. The following twenty eye-tracking and user performance features were used:

   - fixation-related features: fixation-number, mean duration of fixation, standard deviation of fixation duration, maximum fixation duration, minimum fixation duration.
   - saccade-related features: saccade number, mean of blink duration of saccades, mean amplitude of saccades, standard deviation of saccades amplitude, maximum saccade amplitude, minimum saccade amplitude.
   - blink-related features: blink number and mean of blink duration.
   - pupillary response features: mean of left pupil diameter, mean of right pupil diameter, standard deviation of left pupil diameter, standard deviation of right pupil diameter
   - user performance features: related to cognitive DDST test. The features are number of errors, mean response time and response number.

   A binary subject-independent classification was conducted where the first class contained observations with a low level and the second class contained observations with a high level of cognitive workload. The most accurate results were achieved with the linear Support Vector Machine (SVM) – 0.97. The feature selection process allows to obtain better results using a smaller number of features.The cognitive test used in the study is sensitive to changes in cognitive functioning, and correlates with the ability to perform everyday tasks. Eye-tracking features are useful in cognitive workload analysis. This non-invasive method can be applied to measure the level of mental fatigue. The method takes into account memory and concentration of the participant.

Nowadays, the subject-independent approach becomes increasingly more popular. This approach has the potential for predicting cognitive workload because it ensures a high quality of classification, regardless of conditions such as age or habits of the examined person. The model is usually applied to the data that are taken from a new participant. Contrary to the majority of other sources presented in scientific literature, this research does not limit eye-tracking features to pupillary related features but also employs a vast set of other eye-tracking features, including saccades related features, fixation related features, blink-related features and performance related features. This study was one of the first where results of the DSST test were applied for cognitive workload classification.

The presented contributions can be found in an article titled "Binary Classification of Cognitive Workload Levels with Oculography Features" written by Monika Kaczorowska, Martyna Wawrzyk and Małgorzata Plechawska-Wójcik.

2. **Interpretable multiclass subject-independent machine learning model predicting cognitive workload levels.**

This contribution refers to objective 2. I have created an interpretable machine learning model that allows to predict cognitive workload levels using Logistic Regression based on following features: mean amplitude of saccades, standard deviation of fixation duration, fixation number, fixation number, saccade number, mean response time and response number. This model predicts one of the three levels of cognitive workload: low, medium and high. This model has achieved 0.97 of F1 measure using only seven out of twenty features. All of the features were mentioned in the previous section and in the first publication of this thesis.

The interpretable machine learning approach allows to generate a feature ranking and to understand the reasons for the decisions made by a machine learning model. A set of binary classifiers was used to obtain the ranking for each level of cognitive workload. The feature ranking was obtained based on the Logistic Regression model for each class. The ranking can be analytically interpreted using the weight assigned to each feature. The analysis of the importance of processed features using interpretable machine learning techniques allows for a deeper understanding of mental processes. This approach enables the analysis of individual components of the model along with the selection of features that best distinguish each level of cognitive workload.

To the best of my knowledge, no previous work presented an application of interpretable machine learning to cognitive workload estimation. Researchers often performed binary classification instead of multiclass classification. One of the valuable outcomes of the presented approach is to gain the interpretability of features in the process of the subject-independent approach. The results show that the interpretable machine learning approach can enable to obtain better understanding of features as well.

The presented contributions can be found in the article titled "Interpretable Machine Learning Models for three-Way Classification of Cognitive Workload Levels for Eye-Tracking features" written by Monika Kaczorowska, Małgorzata Plechawska-Wójcik and Mikhail Tokovarov.

3. Multiclass subject-independent machine learning model predicting cognitive workload levels using fuzzy aggregation functions.

This contribution refers to objective 3. I have created a machine learning model using the Choquet fuzzy aggregation function that allows to predict cognitive workload levels based on all features mentioned in the first publication of this thesis. The model based on classifier ensembles has achieved higher accuracy levels than separate classifiers. The research in general shows a promising perspective: improvement of classification performance with appropriate combination of individual model results. The model predicts one of the three levels of cognitive workload: low, medium and high.

Aggregation functions allow to improve the classification model by applying the knowledge cumulated in the parameters of the model. This approach is based on applying the probabilities of belonging to each cognitive workload class from original multiclass classification which are the inputs of the aggregation functions.

The literature review shows that using fuzzy aggregation methods for cognitive workload level classification is a new approach. The fuzzy aggregation methods made it possible to obtain better results of classification. The results prove that the use of the generalized Choquet integral method ensures improvement even if the initial individual classifiers provide weak results.

The presented contributions can be found in an article titled "On the Improvement of Eye Tracking-Based Cognitive Workload Estimation Using Aggregation Functions" written by Monika Kaczorowska, Paweł Karczmarek, Małgorzata Plechawska-Wójcik and Mikhail Tokovarov.

4. **Multiclass subject-independent machine learning model predicting cognitive workload levels based on Ex-Gaussian statistics.**

This contribution refers to objective 3. I have created a classification model based on features related to ex-Gaussian parameters which are proper model describing distributions of eye-tracking features. These 16 features are: amplitude of saccade, number of saccades and saccade duration, number of fixations and fixation duration, and number of blinks, number of correct answers and single trial response time. Among the listed features, the number of saccades, fixations, blinks and correct responses were extracted for the specific time intervals (10s period). Information processing shows variable dynamics and temporal changes in its efficiency hence the period of 10s has been chosen for some features.

The Random Forest model has achieved the best result - almost 96% of accuracy and F1 measure based on all features in the multi-class subject-independent classification of cognitive workload. The model predicts one of the three levels of cognitive workload: low, medium and high.

The literature review shows that authors did not use the features calculated based on Ex-Gaussian statistics in the classification process of cognitive workload. The ex-Gaussian statistics describes eye-tracking related features. Features based on ex-Gaussian characteristics associated with cognitive workload data can be used as the input to a classification model. The advantage of using the ex-Gaussian distribution is not only the possibility of obtaining very accurate results but also taking into account the distributional features. These features are usually deleted during analysis according to a parametric approach. However, they turn out to be the most distinguishing in case of the levels of the cognitive workload process. As for eye tracking features with the highest classification powers, the majority cover the tau parameter, for example related to saccadic features.

Additionally, other classifiers were tested and using the Logistic Regression and SVM the feature ranking was obtained. It allows to analyze the features in terms of psychological and cognitive data together and separately. The presented ranking is based on interpretable machine learning and unsupervised learning with clustering. The appliance of K-means algorithm allows to obtain the set of the most valuable features based on feature weights.

The presented contributions can be found in a scientific paper titled "Automated Classification of Cognitive Workload Levels Based on Psychophysiological and Behavioural Variables of Ex-Gaussian Distributional Features" written by Monika Kaczorowska, Małgorzata Plechawska-Wójcik, Mikhail Tokovarov and Paweł Krukow.

3.2 Contributions to Cognitive Science

Cognitive factor analysis based on interpretable machine learning models

This contribution refers to objective 2. My studies suggest further directions regarding continuation research allowing for more precise recognition of the phenomena and changes during the cognitive process. My research might provide new insights into understanding cognitive factors analysis and the dependence between eye-tracking and cognitive features.

Interpretable machine learning classification could help to understand and try to explain the mental fatigue in the context of the cognitive process. It allows to obtain the ranking of features and conduct deep analysis of cognitive features. Separate feature ranking for each level of cognitive workload allowed to define the most valuable features for each level of cognitive workload. In my hypothesis I state that the set of the most valuable features allowing to distinguish between low and high levels include: mean amplitude of saccades, mean response time, standard deviation of fixation duration, response number, fixation number, saccade number.

It also might allow to develop more effective monitoring of cognitive workload in the learning process as well as in the case or neuropsychiatric diseases.

The presented contributions can be found in a scientific paper titled “Interpretable Machine Learning Models for three-Way Classification of Cognitive Workload Levels for Eye-Tracking features” written by Monika Kaczorowska, Małgorzata Plechawska-Wójcik and Mikhail Tokovarov.

4 Conclusion

The measurement of cognitive workload levels is an essential part of cognitive science. The application of machine learning algorithms allows to automate the classification process of the cognitive load. The three main objectives in section 1 were defined and achieved: investigating whether the features based on eye-tracking and user performance can be used to classify cognitive workload levels, development of an interpretable machine learning model that allows classification of cognitive workload levels, as well as improvement of the quality of cognitive workload level classification. In the presented studies, the subject-independent approach was applied and both binary and multiclass classification of cognitive workload levels was performed. The results are promising and enable to understand which features are the most informative in the examined process. Understanding the decisions made by machine learning algorithms is more profitable and valuable than treating them as a black box. Introducing the Ex-Gaussian statistics was beneficial in the research into the multiclass classification of cognitive workload levels. The application of aggregation operators is useful, especially if the basic feature selection is applied. The proposed models could be successfully used in practice, among others, through obtaining high classification accuracy.

Nevertheless, the results described in this thesis have limitations. The level of cognitive workload was fixed for all participants. A preliminary study was used to define the cognitive workload levels: low, medium and high. The medium level was assigned to the original version of the DSST test containing nine symbols and lasting 90s. The low and high levels of cognitive workload were defined based on a pilot study in the form of a participant interview. There exist other methods that allow to evaluate cognitive workload levels for each participant individually. For example, the NASA-TLX scale allows to assign a proper cognitive workload level depending on an individual's cognitive ability. Moreover, it is unclear whether an individual scale of cognitive workload could be applied to train a classification model that is subject-independent. However, taking into consideration the level of difficulty for each participant may provide new insights into the results. Additionally, it is worth noting that the differences in education, age and performed work can have an impact on the analysis. In the presented research, a homogeneous group of participants took part in the experiment. The group consisted of students from technical specialties.

The aspects, which were mentioned as limitations in the last paragraph, will be considered as a starting point in future works. The NASA-TLX scale will be applied to assess the cognitive workload difficulty level. The creation of an original questionnaire adjusted to performed tasks is planned as well. Representatives from various professional groups will be asked to take part in the experiment in order to make the examined group more diverse. Moreover, two additional classification approaches will be tested. The first one will include a combination of the interpretable machine learning approach along with aggregation operators. The second approach is based on applying neural networks and their interpretability aspect which is described in the newest scientific papers.

To sum up, this dissertation can be useful to researchers looking for ideas and techniques to study cognitive workload. This dissertation includes a literature overview concerning cognitive workload, eye-tracking and classification problems.

5 References

1. Qi, P., Ru, H., Gao, L., Zhang, X., Zhou, T., Tian, Y., ... & Sun, Y. (2019). Neural mechanisms of mental fatigue revisited: New insights from the brain connectome. *Engineering*, *5*(2), 276-286.
2. Gevins, A., Smith, M. E., McEvoy, L., & Yu, D. (1997). High-resolution EEG mapping of cortical activation related to working memory: effects of task difficulty, type of processing, and practice. *Cerebral cortex (New York, NY: 1991)*, *7*(4), 374-385.
3. DeLuca, J. (2005). Fatigue, cognition, and mental effort. *Fatigue as a window to the brain*, *37*.
4. Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical issues in ergonomics science*, *3*(2), 159-177.
5. Gopher, D., & Donchin, E. (1986). Workload: An examination of the concept.
6. Miyake, S. (2001). Multivariate workload evaluation combining physiological and subjective measures. *International journal of psychophysiology*, *40*(3), 233-238.
7. Stojmenova, K., & Sodnik, J. (2015). Methods for assessment of cognitive workload in driving tasks. In *ICIST 2015 5th International Conference on Information Society and Technology* (pp. 229-234).
8. Grier, R. A., Warm, J. S., Dember, W. N., Matthews, G., Galinsky, T. L., Szalma, J. L., & Parasuraman, R. (2003). The vigilance decrement reflects limitations in effortful attention, not mindlessness. *Human factors*, *45*(3), 349-359.
9. Galy, E., Cariou, M., & Mélan, C. (2012). What is the relationship between mental workload factors and cognitive load types?. *International Journal of Psychophysiology*, *83*(3), 269-275.
10. Monod, H., Kapitaniak, B., 1999. Ergonomie. Masson, Paris.
11. Sciaraffa, N., Aricò, P., Borghini, G., Flumeri, G. D., Florio, A. D., & Babiloni, F. (2019, November). On the use of machine learning for EEG-based Workload assessment: Algorithms comparison in a realistic task. In *International Symposium on Human Mental Workload: Models and Applications* (pp. 170-185). Springer, Cham.
12. Patten, C. J., Kircher, A., Östlund, J., Nilsson, L., & Svenson, O. (2006). Driver experience and cognitive workload in different traffic environments. *Accident Analysis & Prevention*, *38*(5), 887-894.

13. Khanganba, S. P., & Najar, S. A. (2022). Experience of Cognitive Workload During In-Vehicle Distractions. In *International Conference of the Indian Society of Ergonomics* (pp. 1471-1479). Springer, Cham.
14. Dehais, F., Duprès, A., Blum, S., Drougard, N., Scannella, S., Roy, R. N., & Lotte, F. (2019). Monitoring pilot's mental workload using ERPs and spectral power with a six-dry-electrode EEG system in real flight conditions. *Sensors*, *19*(6), 1324.
15. Antoine, M., Abdessalem, H. B., & Frasson, C. (2022). Cognitive Workload Assessment of Aircraft Pilots. *Journal of Behavioral and Brain Science*, *12*(10), 474-484.
16. Aricò, P., Borghini, G., Di Flumeri, G., Colosimo, A., Pozzi, S., & Babiloni, F. (2016). A passive brain–computer interface application for the mental workload assessment on professional air traffic controllers during realistic air traffic control tasks. *Progress in brain research*, *228*, 295-328.
17. Mazher, M., Abd Aziz, A., Malik, A. S., & Amin, H. U. (2017). An EEG-based cognitive load assessment in multimedia learning using feature extraction and partial directed coherence. *IEEE Access*, *5*, 14819-14829.
18. Zhang, L., Wade, J., Bian, D., Fan, J., Swanson, A., Weitlauf, A., ... & Sarkar, N. (2017). Cognitive load measurement in a virtual reality-based driving system for autism intervention. *IEEE transactions on affective computing*, *8*(2), 176-189.
19. Mathan S, Smart A, Ververs T, Feuerstein M. Towards an index of cognitive efficacy EEG-based estimation of cognitive load among individuals experiencing cancer-related cognitive decline. Annu Int Conf IEEE Eng Med Biol Soc. 2010;2010:6595-8. doi: 10.1109/IEMBS.2010.5627126. PMID: 21096515.
20. Ortega-Morán, J. F., Pagador, J. B., Luis-del-Campo, V., Gómez-Blanco, J. C., & Sánchez-Margallo, F. M. (2019, September). Using eye tracking to analyze surgeons' cognitive workload during an advanced laparoscopic procedure. In *Mediterranean Conference on Medical and Biological Engineering and Computing* (pp. 3-12). Springer, Cham.
21. Wilbanks, B. A., & McMullan, S. P. (2018). A review of measuring the cognitive workload of electronic health records. *CIN: Computers, Informatics, Nursing*, *36*(12), 579-588.
22. Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology* (Vol. 52, pp. 139-183). North-Holland.

23. Reid, G. B., & Nygren, T. E. (1988). The subjective workload assessment technique: A scaling procedure for measuring mental workload. In *Advances in psychology* (Vol. 52, pp. 185-218). North-Holland.
24. Zijlstra, F. R. H., & Van Doorn, L. (1985). *The construction of a scale to measure perceived effort*. University of Technology.
25. Marquart, G., Cabrall, C., & de Winter, J. (2015). Review of eye-related measures of drivers' mental workload. *Procedia Manufacturing*, *3*, 2854-2861.
26. Miller, S. (2001). Workload measures. *National Advanced Driving Simulator. Iowa City, United States*.
27. von Janczewski, N., Kraus, J., Engeln, A., & Baumann, M. (2022). A subjective one-item measure based on NASA-TLX to assess cognitive workload in driver-vehicle interaction. *Transportation research part F: traffic psychology and behaviour*, *86*, 210-225.
28. Zulfany, A. H., Dewi, R. S., & Partiwi, S. G. (2019). Analyzing Mental Workload of Remote Worker by Using SWAT Methodology (Case Study: Remote Software Engineer). IOP Conference Series: Materials Science and Engineering, 598, 012008. doi:10.1088/1757-899x/598/1/01200
29. Zak, Y., Parmet, Y., & Oron-Gilad, T. (2020, October). Subjective Workload assessment technique (SWAT) in real time: affordable methodology to continuously assess human operators' workload. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 2687-2694). IEEE.
30. Ghanbary Sartang, A., Ashnagar, M., Habibi, E., & Sadeghi, S. (2016). Evaluation of Rating Scale Mental Effort (RSME) effectiveness for mental workload assessment in nurses. *Journal of Occupational Health and Epidemiology*, *5*(4), 211-217.
31. Tattersall, A. J., & Foord, P. S. (1996). An experimental evaluation of instantaneous self-assessment as a measure of workload. Ergonomics, 39(5), 740-748.
32. Bennett, S. A. (2018). Pilot workload and fatigue on short-haul routes: an evaluation supported by instantaneous self-assessment and ethnography. *Journal of Risk Research*, *21*(5), 645-677.
33. Rubio, S., Díaz, E., Martín, J., & Puente, J. M. (2004). Evaluation of subjective mental workload: A comparison of SWAT, NASA-TLX, and workload profile methods. *Applied psychology*, *53*(1), 61-86.

34. Joseph, A. W., Vaiz, J. S., & Murugesh, R. (2021, July). Modeling Cognitive Load in Mobile Human Computer Interaction Using Eye Tracking Metrics. In *International Conference on Applied Human Factors and Ergonomics* (pp. 99-106). Springer, Cham.

35. Bruyer, R., & Brysbaert, M. (2011). 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)?. *Psychologica Belgica*, *51*(1), 5-13.

36. Zarjam, P., Epps, J., Chen, F., & Lovell, N. H. (2013). Estimating cognitive workload using wavelet entropy-based features during an arithmetic task. *Computers in biology and medicine*, *43*(12), 2186-2195.

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

38. Harel, J., Koch, C., & Perona, P. (2006). Graph-based visual saliency. *Advances in neural information processing systems*, *19*.

39. Ktistakis, E., Skaramagkas, V., Manousos, D., Tachos, N. S., Tripoliti, E., Fotiadis, D. I., & Tsiknakis, M. (2022). COLET: A dataset for COgnitive workLoad estimation based on eye-tracking. *Computer Methods and Programs in Biomedicine*, *224*, 106989.

40. Knoll, A., Wang, Y., Chen, F., Xu, J., Ruiz, N., Epps, J., & Zarjam, P. (2011, September). Measuring cognitive workload with low-cost electroencephalograph. In *Ifip conference on human-computer interaction* (pp. 568-571). Springer, Berlin, Heidelberg.

41. Ziegler, M. D., Kraft, A., Krein, M., Lo, L. C., Hatfield, B., Casebeer, W., & Russell, B. (2016, July). Sensing and assessing cognitive workload across multiple tasks. In *International Conference on Augmented Cognition* (pp. 440-450). Springer, Cham.

42. Hajinoroozi, M., Mao, Z., Jung, T. P., Lin, C. T., & Huang, Y. (2016). EEG-based prediction of driver's cognitive performance by deep convolutional neural network. *Signal Processing: Image Communication*, *47*, 549-555.

43. Gamboa, H., Silva, H., & Fred, A. (2014). HiMotion: a new research resource for the study of behavior, cognition, and emotion. Multimedia tools and applications, 73, 345-375.

44. Dimitrakopoulos, G. N., Kakkos, I., Dai, Z., Lim, J., deSouza, J. J., Bezerianos, A., & Sun, Y. (2017). Task-independent mental workload classification based upon common multiband EEG cortical connectivity. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *25*(11), 1940-1949.

45. Wang, S., Gwizdka, J., & Chaovalitwongse, W. A. (2015). Using wireless EEG signals to assess memory workload in the $ n $-back task. *IEEE Transactions on Human-Machine Systems, 46*(3), 424-435.

46. Tremmel, C., Herff, C., Sato, T., Rechowicz, K., Yamani, Y., & Krusienski, D. J. (2019). Estimating cognitive workload in an interactive virtual reality environment using EEG. *Frontiers in human neuroscience*, *13*, 401.

47. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. *In Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2921-2929).

48. Cook, D. B., O'Connor, P. J., Lange, G., & Steffener, J. (2007). Functional neuroimaging correlates of mental fatigue induced by cognition among chronic fatigue syndrome patients and controls. *Neuroimage*, *36*(1), 108-122.

49. Comstock Jr, J. R., & Arnegard, R. J. (1992). *The multi-attribute task battery for human operator workload and strategic behavior research* (No. NAS 1.15: 104174).

50. Chandra, S., Sharma, G., Verma, K. L., Mittal, A., & Jha, D. (2015). EEG based cognitive workload classification during NASA MATB-II multitasking. International Journal of Cognitive Research in Science, Engineering and Education, 3(1), 35-42.

51. Pradhan, G. N., Hagen, K. M., Cevette, M. J., & Stepanek, J. (2022, June). Oculo-Cognitive Addition Test: Quantifying Cognitive Performance During Variable Cognitive Workload Through Eye Movement Features. In *2022 IEEE 10th International Conference on Healthcare Informatics (ICHI)* (pp. 422-430). IEEE.

52. Boake, C. (2002). From the Binet–Simon to the Wechsler–Bellevue: Tracing the history of intelligence testing. *Journal of clinical and experimental neuropsychology*, 24(3), 383-405.

53. Jaeger, J. (2018). Digit symbol substitution test: the case for sensitivity over specificity in neuropsychological testing. *Journal of clinical psychopharmacology*, *38*(5), 513.

54. Sicard, V., Moore, R. D., & Ellemberg, D. (2019). Sensitivity of the Cogstate test battery for detecting prolonged cognitive alterations stemming from sport-related concussions. *Clinical journal of sport medicine*, *29*(1), 62-68.

55. Brunken, R., Plass, J. L., & Leutner, D. (2003). Direct measurement of cognitive load in multimedia learning. *Educational psychologist, 38(1),* 53-61.

56. Gentili, R. J., Rietschel, J. C., Jaquess, K. J., Lo, L. C., Prevost, C. M., Miller, M. W., ... & Hatfield, B. D. (2014, August). Brain biomarkers based assessment of cognitive workload in pilots under various task demands. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 5860-5863). IEEE.

57. Ayaz, H., Willems, B., Bunce, B., Shewokis, P. A., Izzetoglu, K., Hah, S., ... & Onaral, B. (2010). Cognitive workload assessment of air traffic controllers using optical brain imaging sensors. *Advances in understanding human performance: Neuroergonomics, human factors design, and special populations,* 21-31.

58. Bhavsar, P. (2022). Context-dependent cognitive workload monitoring using pupillometry for control room operators to prevent overload. *IISE transactions on occupational ergonomics and human factors,* (just-accepted), 1-16.

59. Wobrock, D., Frey, J., Graeff, D., Rivière, J. B. D. L., Castet, J., & Lotte, F. (2015, September). Continuous mental effort evaluation during 3d object manipulation tasks based on brain and physiological signals. In *IFIP Conference on Human-Computer Interaction* (pp. 472-487). Springer, Cham.

60. Matthews, G., Reinerman-Jones, L. E., Barber, D. J., & Abich IV, J. (2015). The psychometrics of mental workload: Multiple measures are sensitive but divergent. *Human factors*, *57*(1), 125-143.

61. Mehler, B., Reimer, B., & Wang, Y. (2011, June). A comparison of heart rate and heart rate variability indices in distinguishing single-task driving and driving under secondary cognitive workload. *In Driving Assesment Conference* (Vol. 6, No. 2011). University of Iowa.

62. Elul, R. (1972). The genesis of the EEG. *International review of neurobiology*, *15*, 227-272.

63. Quaresima, V., & Ferrari, M. (2019, August). A mini-review on functional near-infrared spectroscopy (fNIRS): where do we stand, and where should we go?. In *Photonics* (Vol. 6, No. 3, p. 87). MDPI.

64. Zhou, Y., Huang, S., Xu, Z., Wang, P., Wu, X., & Zhang, D. (2021). Cognitive workload recognition using EEG signals and machine learning: A review. *IEEE Transactions on Cognitive and Developmental Systems*.

65. Belkhiria, C., & Peysakhovich, V. (2021). EOG metrics for cognitive workload detection. *Procedia Computer Science*, *192*, 1875-1884.

66. Čegovnik, T., Stojmenova, K., Jakus, G., & Sodnik, J. (2018). An analysis of the suitability of a low-cost eye tracker for assessing the cognitive load of drivers. *Applied ergonomics*, *68*, 1-11.

67. Duchowski, A. T., Krejtz, K., Krejtz, I., Biele, C., Niedzielska, A., Kiefer, P., ... & Giannopoulos, I. (2018, April). The index of pupillary activity: Measuring cognitive load

vis-à-vis task difficulty with pupil oscillation. *In Proceedings of the 2018 CHI conference on human factors in computing systems* (pp. 1-13).

68. Ye, Y., Shi, Y., Xia, P., Kang, J., Tyagi, O., Mehta, R. K., & Du, J. (2022). Cognitive characteristics in firefighter wayfinding Tasks: An Eye-Tracking analysis. *Advanced Engineering Informatics*, *53*, 101668.
69. Richardson, J. T. (2007). Measures of short-term memory: a historical review. *Cortex*, 43(5), 635-650.
70. Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1), 49-59.
71. Venables, P. H., & Christie, M. J. (1980). Electrodermal activity. *Techniques in psychophysiology*, *54*(3).
72. Nourbakhsh, N., Wang, Y., Chen, F., & Calvo, R. A. (2012, November). Using galvanic skin response for cognitive load measurement in arithmetic and reading tasks. In *Proceedings of the 24th australian computer-human interaction conference* (pp. 420-423).
73. Widyanti, A., Muslim, K., & Sutalaksana, I. Z. (2017). The sensitivity of Galvanic Skin Response for assessing mental workload in Indonesia. *Work*, *56*(1), 111-117.
74. Hughes, A. M., Hancock, G. M., Marlow, S. L., Stowers, K., & Salas, E. (2019). Cardiac measures of cognitive workload: a meta-analysis. *Human factors*, *61*(3), 393-414.
75. Nourbakhsh, N., Wang, Y., & Chen, F. (2013, September). GSR and blink features for cognitive load classification. In *IFIP conference on human-computer interaction* (pp. 159-166). Springer, Berlin, Heidelberg.
76. Mark, J., Curtin, A., Kraft, A., Sargent, A., Perez, A., Friedman, L., ... & Ayaz, H. (2018). Multimodal Cognitive Workload Assessment Using EEG, fNIRS, ECG, EOG, PPG, and Eye-tracking. *Frontiers in Human Neuroscience*, *12*.
77. Borghini, G., Aricò, P., Graziani, I., Salinari, S., Sun, Y., Taya, F., ... & Babiloni, F. (2016). Quantitative assessment of the training improvement in a motor-cognitive task by using EEG, ECG and EOG signals. *Brain topography*, *29*(1), 149-161.
78. Othman, N., & Romli, F. I. (2016). Mental workload evaluation of pilots using pupil dilation. *International Review of Aerospace Engineering*, *9*(3), 80-84.

79. Fritz, T., Begel, A., Müller, S. C., Yigit-Elliott, S., & Züger, M. (2014, May). Using psycho-physiological measures to assess task difficulty in software development. In *Proceedings of the 36th international conference on software engineering* (pp. 402-413).
80. Broadbent, D. P., D'Innocenzo, G., Ellmers, T. J., Parsler, J., Szameitat, A. J., & Bishop, D. T. (2023). Cognitive load, working memory capacity and driving performance: A preliminary fNIRS and eye tracking study. *Transportation Research Part F: Traffic Psychology and Behaviour*, *92*, 121-132.
81. Wade, N., & Tatler, B. W. (2005). *The moving tablet of the eye: The origins of modern eye movement research*. Oxford University Press, USA.
82. Bell, C. (1823). XV. On the motions of the eye, in illustration of the uses of the muscles and nerves of the orbit. *Philosophical Transactions of the Royal Society of London*, (113), 166-186.
83. Carter, B. T., & Luke, S. G. (2020). Best practices in eye tracking research. *International Journal of Psychophysiology*, *155*, 49-62.
84. Clarke, A. D., Mahon, A., Irvine, A., & Hunt, A. R. (2017). People are unable to recognize or report on their own eye movements. *The Quarterly Journal of Experimental Psychology*, *70*(11), 2251-2270.
85. Korbach, A., Ginns, P., Brünken, R., & Park, B. (2020). Should learners use their hands for learning? Results from an eye-tracking study. *Journal of Computer Assisted Learning*, *36*(1), 102-113.
86. Harezlak, K., & Kasprowski, P. (2018). Application of eye tracking in medicine: A survey, research issues and challenges. *Computerized Medical Imaging and Graphics*, *65*, 176-190.
87. Wan, G., Kong, X., Sun, B., Yu, S., Tu, Y., Park, J., ... & Kong, J. (2019). Applying eye tracking to identify autism spectrum disorder in children. *Journal of autism and developmental disorders*, *49*(1), 209-215.
88. Harezlak, K., & Kasprowski, P. (2018). Application of eye tracking in medicine: A survey, research issues and challenges. *Computerized Medical Imaging and Graphics*, *65*, 176-190.
89. Borkowska, A. R., & Francuz, P. (2013). Ruchy gałek ocznych podczas oceny poprawności zapisu wyrazów jako wskaźnik rozwoju świadomości ortograficznej młodzieży z dysortografią. *Psychologia rozwojowa, 18(3).*
90. Rybakowski, J. K., Borkowska, A., Czerski, P. M., & Hauser, J. (2001). Dopamine D3 receptor (DRD3) gene polymorphism is associated with the intensity of eye movement

disturbances in schizophrenic patients and healthy subjects. *Molecular psychiatry*, 6(6), 718-724.

91. Bojko, A. (2013). *Eye tracking the user experience: A practical guide to research*. Rosenfeld Media.
92. Duerrschmid, K., & Danner, L. (2018). Eye tracking in consumer research. In *Methods in Consumer Research, Volume 2* (pp. 279-318). Woodhead Publishing.
93. Peißl, S., Wickens, C. D., & Baruah, R. (2018). Eye-tracking measures in aviation: A selective literature review. *The International Journal of Aerospace Psychology*, *28*(3-4), 98-112.
94. Xu, J., Min, J., & Hu, J. (2018). Real-time eye tracking for the assessment of driver fatigue. *Healthcare technology letters*, *5*(2), 54-58.
95. Liao, H., Zhao, W., Zhang, C., Dong, W., & Huang, H. (2022). Detecting individuals' spatial familiarity with urban environments using eye movement data. *Computers, Environment and Urban Systems*, *93*, 101758.
96. Begel, A., & Vrzakova, H. (2018, June). Eye movements in code review. In *Proceedings of the Workshop on Eye Movements in Programming* (pp. 1-5).
97. Sharafi, Z., Sharif, B., Guéhéneuc, Y. G., Begel, A., Bednarik, R., & Crosby, M. (2020). A practical guide on conducting eye tracking studies in software engineering. *Empirical Software Engineering, 25(5),* 3128-3174.
98. Carizio, B. G., Silva, G. A., Paschoalino, G. P., de Angelo, J. C., Gotardi, G. C., Polastri, P. F., & Rodrigues, S. T. (2021). Pupil dilation as indicative of cognitive workload while driving a car: effects of cell phone use and driver experience in young adults. *Brazilian Journal of Motor Behavior*, *15*(5), 391-402.
99. Martinez-Marquez, D., Pingali, S., Panuwatwanich, K., Stewart, R. A., & Mohamed, S. (2021). Application of eye tracking technology in aviation, maritime, and construction industries: a systematic review. *Sensors, 21(13),* 4289.
100. Mallick, R., Slayback, D., Touryan, J., Ries, A. J., & Lance, B. J. (2016, October). The use of eye metrics to index cognitive workload in video games. In *2016 IEEE Second Workshop on Eye Tracking and Visualization (ETVIS)* (pp. 60-64). IEEE.
101. Meena, Y. K., Cecotti, H., Wong-Lin, K., & Prasad, G. (2017, July). A multimodal interface to resolve the Midas-Touch problem in gaze controlled wheelchair. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 905-908). IEEE.

102. Henderson, J. M., Shinkareva, S. V., Wang, J., Luke, S. G., & Olejarczyk, J. (2013). Predicting cognitive state from eye movements. *PloS one, 8*(5), e64937.

103. Nilsson Benfatto, M., Öqvist Seimyr, G., Ygge, J., Pansell, T., Rydberg, A., & Jacobson, C. (2016). Screening for dyslexia using eye tracking during reading. *PloS one, 11*(12), e0165508.

104. Yamada, Y., & Kobayashi, M. (2018). Detecting mental fatigue from eye-tracking data gathered while watching video: Evaluation in younger and older adults. *Artificial intelligence in medicine, 91*, 39-48.

105. Rello, L., & Ballesteros, M. (2015, May). Detecting readers with dyslexia using machine learning with eye tracking measures. *In Proceedings of the 12th International Web for All Conference* (pp. 1-8).

106. Moon, S., Kahya, M., Lyons, K. E., Pahwa, R., Akinwuntan, A. E., & Devos, H. (2021). Cognitive workload during verbal abstract reasoning in Parkinson's disease: Apilot study. *International Journal of Neuroscience, 131*(5), 504-510.

107. Tolvanen, O., Elomaa, A. P., Itkonen, M., Vrzakova, H., Bednarik, R., & Huotarinen, A. (2022). Eye-Tracking Indicators of Workload in Surgery: A Systematic Review. *Journal of Investigative Surgery, 35*(6), 1340-1349.

108. Gil, A. M., Birdi, S., Kishibe, T., & Grantcharov, T. P. (2022). Eye Tracking Use in Surgical Research: A Systematic Review. *Journal of Surgical Research, 279*, 774-787.

109. Wisiecka, K., Krejtz, K., Krejtz, I., Sromek, D., Cellary, A., Lewandowska, B., & Duchowski, A. (2022, June). Comparison of Webcam and Remote Eye Tracking. In 2022 Symposium on Eye Tracking Research and Applications (pp. 1-7).

110. Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological bulletin, 124*(3), 372.

111. Rayner, K. (2009). The 35th Sir Frederick Bartlett Lecture: Eye movements and attention in reading, scene perception, and visual search. *Quarterly journal of experimental psychology, 62*(8), 1457-1506.

112. Duchowski, A. T., & Duchowski, A. T. (2017). *Eye tracking methodology: Theory and practice*. Springer.

113. Mathôt, S., Fabius, J., Van Heusden, E., & Van der Stigchel, S. (2018). Safe and sensible preprocessing and baseline correction of pupil-size data. *Behavior research methods, 50*(1), 94-106.

114. Mathôt, S., & Vilotijević, A. (2022). Methods in Cognitive Pupillometry: Design, *Preprocessing, and Statistical Analysis*. bioRxiv.

115. Hollander, J., & Huette, S. (2022). Extracting blinks from continuous eye-tracking data in a mind wandering paradigm. *Consciousness and Cognition*, *100*, 103303.

116. Pedrotti, M., Lei, S., Dzaack, J., & Rötting, M. (2011). A data-driven algorithm for offline pupil signal preprocessing and eyeblink detection in low-speed eye-tracking protocols. *Behavior Research Methods*, *43*(2), 372-383.

117. Joseph, A. W., & Murugesh, R. (2020). Potential eye tracking metrics and indicators to measure cognitive load in human-computer interaction research. *J. Sci. Res*, *64*(1), 168-175.

118. Krejtz, K., Duchowski, A. T., Niedzielska, A., Biele, C., & Krejtz, I. (2018). Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze. *PloS one, 13(9),* e0203629.

119. Hessels, R. S., Niehorster, D. C., Nyström, M., Andersson, R., & Hooge, I. T. (2018). Is the eye-movement field confused about fixations and saccades? A survey among 124 researchers. *Royal Society open science*, *5*(8), 180502.

120. Kacur, J., Polec, J., & Csóka, F. (2019, September). Eye tracking and KNN based detection of schizophrenia. *In 2019 International Symposium ELMAR* (pp. 123-126). IEEE.

121. Shishido, E., Ogawa, S., Miyata, S., Yamamoto, M., Inada, T., & Ozaki, N. (2019). Application of eye trackers for understanding mental disorders: Cases for schizophrenia and autism spectrum disorder. *Neuropsychopharmacology reports*, *39*(2), 72-77.

122. Kardan, O., Berman, M. G., Yourganov, G., Schmidt, J., & Henderson, J. M. (2015). Classifying mental states from eye movements during scene viewing. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(6), 1502.

123. Dowiasch, S., Marx, S., Einhäuser, W., & Bremmer, F. (2015). Effects of aging on eye movements in the real world. *Frontiers in human neuroscience*, *9*, 46.

124. Li, F., Xu, G., & Feng, S. (2021, October). Eye Tracking Analytics for Mental States Assessment–A Review. In *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 2266-2271). IEEE.

125. Skaramagkas, V., Giannakakis, G., Ktistakis, E., Manousos, D., Karatzanis, I., Tachos, N., ... & Tsiknakis, M. (2021). Review of eye tracking metrics involved in emotional and cognitive processes. *IEEE Reviews in Biomedical Engineering*.

126. Guo, Y., Freer, D., Deligianni, F., & Yang, G. Z. (2021). Eye-tracking for performance evaluation and workload estimation in space telerobotic training. *IEEE Transactions on Human-Machine Systems*, *52*(1), 1-11.

127. Lacouture, Y., & Cousineau, D. (2008). How to use MATLAB to fit the ex-Gaussian and other probability functions to a distribution of response times. *Tutorials in quantitative methods for psychology*, *4*(1), 35-45.

128. Otero-Millan, J., Troncoso, X. G., Macknik, S. L., Serrano-Pedraza, I., & Martinez-Conde, S. (2008). Saccades and microsaccades during visual fixation, exploration, and search: foundations for a common saccadic generator. *Journal of vision*, *8*(14), 21-21.

129. Guy, N., Lancry-Dayan, O. C., & Pertzov, Y. (2020). Not all fixations are created equal: The benefits of using ex-Gaussian modeling of fixation durations. *Journal of vision*, *20*(10), 9-9.

130. Luke, S. G., Darowski, E. S., & Gale, S. D. (2018). Predicting eye-movement characteristics across multiple tasks from working memory and executive control. *Memory & Cognition*, *46*(5), 826-839.

131. Karakula-Juchnowicz, H., Gałęcka, M., Rog, J., Bartnicka, A., Łukaszewicz, Z., Krukow, P., ... & Juchnowicz, D. (2018). The food-specific serum IgG reactivity in major depressive disorder patients, irritable bowel syndrome patients and healthy controls. *Nutrients*, *10*(5), 548.

132. Chen, L. L., Zhao, Y., Ye, P. F., Zhang, J., & Zou, J. Z. (2017). Detecting driving stress in physiological signals based on multimodal feature analysis and kernel classifiers. *Expert Systems with Applications*, *85*, 279-291.

133. Almogbel, M. A., Dang, A. H., & Kameyama, W. (2018, February). EEG-signals based cognitive workload detection of vehicle driver using deep learning. In *2018 20th International Conference on Advanced Communication Technology (ICACT)* (pp. 256-259). IEEE.

134. Hefron, R., Borghetti, B., Schubert Kabban, C., Christensen, J., & Estepp, J. (2018). Cross-participant EEG-based assessment of cognitive workload using multi-path convolutional recurrent neural networks. *Sensors*, *18*(5), 1339.

135. Lobo, J. L., Ser, J. D., De Simone, F., Presta, R., Collina, S., & Moravek, Z. (2016, September). Cognitive workload classification using eye-tracking and EEG data. In *Proceedings of the International Conference on Human-Computer Interaction in Aerospace* (pp. 1-8).

136. Wang, Z., Hope, R. M., Wang, Z., Ji, Q., & Gray, W. D. (2012). Cross-subject workload classification with a hierarchical Bayes model. *NeuroImage*, *59*(1), 64-69.

137. McKendrick, R., Feest, B., Harwood, A., & Falcone, B. (2019). Theories and methods for labeling cognitive workload: Classification and transfer learning. *Frontiers in human neuroscience*, *13*, 295.

138. Fazli, S., Mehnert, J., Steinbrink, J., Curio, G., Villringer, A., Müller, K. R., & Blankertz, B. (2012). Enhanced performance by a hybrid NIRS–EEG brain computer interface. *Neuroimage*, *59*(1), 519-529.

139. Spüler, M., Walter, C., Rosenstiel, W., Gerjets, P., Moeller, K., & Klein, E. (2016). EEG-based prediction of cognitive workload induced by arithmetic: a step towards online adaptation in numerical learning. *Zdm*, *48*(3), 267-278.

140. Walter, C., Wolter, P., Rosenstiel, W., Bogdan, M., & Spüler, M. (2014, September). Towards cross-subject workload prediction. In *Proceedings of the 6th International Brain-Computer Interface Conference*.

141. Khushaba, R. N., Kodagoda, S., Lal, S., & Dissanayake, G. (2010). Driver drowsiness classification using fuzzy wavelet-packet-based feature-extraction algorithm. *IEEE transactions on biomedical engineering*, *58*(1), 121-131.

142. Chen, W., Sawaragi, T., & Hiraoka, T. (2022). Comparing eye-tracking metrics of mental workload caused by NDRTs in semi-autonomous driving. *Transportation research part F: traffic psychology and behaviour*, *89*, 109-128.

143. Zarjam, P., Epps, J., & Lovell, N. H. (2015). Beyond subjective self-rating: EEG signal classification of cognitive workload. *IEEE Transactions on Autonomous Mental Development*, *7*(4), 301-310.

144. Maiorana, E. (2020). Deep learning for EEG-based biometric recognition. *Neurocomputing*, *410*, 374-386.

145. Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, *46*(1), 389-422.

146. Andrich, D. (1988). Rasch models for measurement (Vol. 68). Sage.

147. Mander, J. B., Priestley, M. J., & Park, R. (1988). Theoretical stress-strain model for confined concrete. Journal of structural engineering, 114(8), 1804-1826.

148. Ramberg, W., & Osgood, W. R. (1943). Description of stress-strain curves by three parameters (No. NACA-TN-902).

149. Baldwin, C. L. (2003). Neuroergonomics of mental workload: New insights from the convergence of brain and behaviour in ergonomics research. *Theoretical Issues in Ergonomics Science*.

150. He, D., Wang, Z., Khalil, E. B., Donmez, B., Qiao, G., & Kumar, S. (2022). Classification of Driver Cognitive Load: Exploring the Benefits of Fusing Eye-Tracking and Physiological Measures. *Transportation Research Record*, 03611981221090937.

151. Thomas, H. B. G. (1963). Communication theory and the constellation hypothesis of calculation. Quarterly Journal of Experimental Psychology, 15(3), 173-191.

152. İşbilir, E., Çakır, M. P., Acartürk, C., & Tekerek, A. Ş. (2019). Towards a multimodal model of cognitive workload through synchronous optical brain imaging and eye tracking measures. *Frontiers in human neuroscience, 13*, 375.

153. Almogbel, M. A., Dang, A. H., & Kameyama, W. (2019, February). Cognitive workload detection from raw EEG-signals of vehicle driver using deep learning. In *2019 21st International Conference on Advanced Communication Technology (ICACT)* (pp. 1-6). IEEE.

154. Fatimah, B., Pramanick, D., & Shivashankaran, P. (2020, July). Automatic detection of mental arithmetic task and its difficulty level using EEG signals. In *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* (pp. 1-6). IEEE.

155. Chakladar, D. D., Dey, S., Roy, P. P., & Iwamura, M. (2021, January). EEG-based cognitive state assessment using deep ensemble model and filter bank common spatial pattern. In *2020 25th International Conference on Pattern Recognition (ICPR)* (pp. 4107-4114). IEEE.

156. Becerra-Sánchez, P., Reyes-Munoz, A., & Guerrero-Ibañez, A. (2020). Feature selection model based on EEG signals for assessing the cognitive workload in drivers. *Sensors*, *20*(20), 5881.

157. Han, S. Y., Kwak, N. S., Oh, T., & Lee, S. W. (2020). Classification of pilots' mental states using a multimodal deep learning network. *Biocybernetics and Biomedical Engineering, 40*(1), 324-336.

158. Gupta, S. S., Taori, T. J., Ladekar, M. Y., Manthalkar, R. R., Gajre, S. S., & Joshi, Y. V. (2021). Classification of cross task cognitive workload using deep recurrent network with modelling of temporal dynamics. *Biomedical Signal Processing and Control, 70*, 103070.

159. Taori, T. J., Gupta, S. S., Gajre, S. S., & Manthalkar, R. R. (2022). Cognitive workload classification: Towards generalization through innovative pipeline interface using HMM. *Biomedical Signal Processing and Control, 78,* 104010.

160. Zheng, Z., Yin, Z., Wang, Y., & Zhang, J. (2023). Inter-subject cognitive workload estimation based on a cascade ensemble of multilayer autoencoders. *Expert Systems with Applications*, *211*, 118694.

161. Bozkir, E., Geisler, D., & Kasneci, E. (2019, March). Person independent, privacy preserving, and real time assessment of cognitive load using eye tracking in a virtual reality setup. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)* (pp. 1834-1837). IEEE.

162. Tombaugh, T. N. (2006). A comprehensive review of the paced auditory serial addition test (PASAT). *Archives of clinical neuropsychology*, 21(1), 53-76.

163. Appel, T., Scharinger, C., Gerjets, P., & Kasneci, E. (2018, June). Cross-subject workload classification using pupil-related measures. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications* (pp. 1-8).

164. Scharinger, C., Soutschek, A., Schubert, T., & Gerjets, P. (2015). When flanker meets the n-back: What EEG and pupil dilation data reveal about the interplay between the two central-executive working memory functions inhibition and updating. *Psychophysiology, 52*(10), 1293-1304.

165. Marshall, S. P. (2000). *U.S. Patent No. 6,090,051*. Washington, DC: U.S. Patent and Trademark Office.

166. Fridman, L., Reimer, B., Mehler, B., & Freeman, W. T. (2018, April). Cognitive load estimation in the wild. In *Proceedings of the 2018 chi conference on human factors in computing systems* (pp. 1-9).

167. Farha, N. A., Al-Shargie, F., Tariq, U., & Al-Nashash, H. (2021, October). Artifact Removal of Eye Tracking Data for the Assessment of Cognitive Vigilance Levels. In *2021 Sixth International Conference on Advances in Biomedical Engineering (ICABME)* (pp. 175-179). IEEE.

168. Wu, C., Cha, J., Sulek, J., Zhou, T., Sundaram, C. P., Wachs, J., & Yu, D. (2020). Eye-tracking metrics predict perceived workload in robotic surgical skills training. *Human factors, 62*(8), 1365-1386.

169. Bitkina, O. V., Park, J., & Kim, H. K. (2021). The ability of eye-tracking metrics to classify and predict the perceived driving workload. *International Journal of Industrial Ergonomics*, *86,* 103193.

170. Zahabi, M., Wang, Y., & Shahrampour, S. (2021). Classification of Officers' Driving Situations Based on Eye-Tracking and Driver Performance Measures. *IEEE Transactions on Human-Machine Systems, 51*(4), 394-402.

171. Shi, C., Rothrock, L., & Noah, B. (2023). Using Eye-Tracking to Predict Cognitive Workload in a Control Room Environment. In *Human-Automation Interaction* (pp. 217-233). Springer, Cham.

172. Appel, T., Sevcenko, N., Wortha, F., Tsarava, K., Moeller, K., Ninaus, M., ... & Gerjets, P. (2019, October). Predicting cognitive load in an emergency simulation based on behavioral and physiological measures. In *2019 International Conference on Multimodal Interaction* (pp. 154-163).

173. Aygun, A., Lyu, B., Nguyen, T., Haga, Z., Aeron, S., & Scheutz, M. (2022, November). Cognitive Workload Assessment via Eye Gaze and EEG in an Interactive Multi-Modal Driving Task. In *Proceedings of the 2022 International Conference on Multimodal Interaction* (pp. 337-348).

174. Shinde, P. P., & Shah, S. (2018, August). A review of machine learning and deep learning applications. *In 2018 Fourth international conference on computing communication control and automation (ICCUBEA)* (pp. 1-6). IEEE.

175. Bereta, M., Karczmarek, P., Pedrycz, W., & Reformat, M. (2013). Local descriptors in application to the aging problem in face recognition. *Pattern Recognition,* 46(10), 2634-2646.

176. Cpałka, K., Zalasiński, M., & Rutkowski, L. (2014). New method for the on-line signature verification based on horizontal partitioning. *Pattern Recognition, 47(8),* 2652-2661.

177. Caywood, M. S., Roberts, D. M., Colombe, J. B., Greenwald, H. S., & Weiland, M. Z. (2017). Gaussian process regression for predictive but interpretable machine learning models: an example of predicting mental workload across tasks. *Frontiers in human neuroscience, 10,* 647.

178. Cui, J., Lan, Z., Liu, Y., Li, R., Li, F., Sourina, O., & Müller-Wittig, W. (2022). A compact and interpretable convolutional neural network for cross-subject driver drowsiness detection from single-channel EEG. *Methods, 202,* 173-184.

179. Ieracitano, C., Mammone, N., Hussain, A., & Morabito, F. C. (2022). A novel explainable machine learning approach for EEG-based brain-computer interface systems. *Neural Computing and Applications, 34(14),* 11347-11360.

180. Islam, R., Andreev, A. V., Shusharina, N. N., & Hramov, A. E. (2022). Explainable machine learning methods for classification of brain states during visual perception. *Mathematics, 10(15),* 2819.

181. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems, 30*.

182. Grabisch, M., Marichal, J. L., Mesiar, R., & Pap, E. (2009). Aggregation functions (Vol. 127). Cambridge University Press.

183. Choquet, G. (1954). Theory of capacities. In Annales d' l'institut Fourier (Vol. 5, pp. 131-295).

184. Bustince, H., Sanz, J. A., Lucca, G., Dimuro, G. P., Bedregal, B., Mesiar, R., ... & Ochoa, G. (2016, July). Pre-aggregation functions: definition, properties and construction methods. *In 2016 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)* (pp. 294-300). IEEE.

185. Klement, E. P., & Mesiar, R. (Eds.). (2005). Logical, algebraic, analytic and probabilistic aspects of triangular norms. Elsevier.

186. Karczmarek, P., Kiersztyn, A., & Pedrycz, W. (2018). Generalized Choquet integral for face recognition. *International Journal of Fuzzy Systems, 20,* 1047-1055.

187. Karczmarek, P. (2018). Selected problems of face recognition and decision-making theory. Wydawnictwo Politechniki Lubelskiej.

6 Scientific papers comprising the thesis

6.1 Binary Classification of Cognitive Workload Levels with Oculography Features

1. Details

This article was written by Monika Kaczorowska, Martyna Warzyk, and Małgorzata Plechawska-Wójcik and published in Proceedings of Computer Information Systems and Industrial Management: 19th International Conference, CISIM 2020 Bialystok, Poland, October 16–18, 2020. The article is worth 40 points according to the list of the Polish Ministry of Science and Higher Education.

2. Abstract

Assessment of cognitive workload level is important to understand human mental fatigue, especially in the case of performing intellectual tasks. The paper presents a case study regarding binary classification of cognitive workload levels. The dataset was received from two versions of the digit symbol substitution test (DSST), conducted on 26 healthy volunteers. A screen-based eye tracker was applied during the examination gathering oculographic data. The DSST tests results such as total number of matches and error ratio were also applied. The classification was performed with several different machine learning models. The best accuracy (97%) was achieved with the linear SVM classifier. The final dataset for classification was based on nine features selected using the Fisher score feature selection method.

Binary Classification of Cognitive Workload Levels with Oculography Features

Monika Kaczorowska, Martyna Wawrzyk,
and Małgorzata Plechawska-Wójcik(✉)

Lublin University of Technology, Nadbystrzycka 36B, 20-618 Lublin, Poland
{m.kaczorowska,m.plechawska}@pollub.pl,
martyna.wawrzyk@pollub.edu.pl

Abstract. Assessment of cognitive workload level is important to understand human mental fatigue, especially in the case of performing intellectual tasks. The paper presents a case study on binary classification of cognitive workload levels. The dataset was received from two versions of the digit symbol substitution test (DSST), conducted on 26 healthy volunteers. A screen-based eye tracker was applied during an examination gathering oculographic data. DSST test results such as total number of matches and error ratio were also applied. Classification was performed with several different machine learning models. The best accuracy (97%) was achieved with linear SVM classifier. The final dataset for classification was based on nine features selected with the Fisher score feature selection method.

Keywords: Cognitive workload · Binary classification · SVM · Eye-tracking signal

1 Introduction

According to the literature, the term "cognitive workload" is a quantitative measure of the amount of mental effort necessary to perform a task [1]. Estimation of cognitive workload is of great importance in understanding human mental fatigue related to performing tasks of various complexity requiring different concentration level. Moreover, assessment of mental effort might be useful in the process of modeling information processing capabilities.

The Digit Symbol Substitution Test (DSST) [2, 3], known from over a century ago, was introduced as a tool to understand human associative learning. Currently it is one of the most commonly used tests in clinical neuropsychology to measure cognitive dysfunction. Its popularity is related to its brevity and high discriminant validity [4]. The DSST enables to check the processing speed, memory and executive functioning of the patient. It is prevalent in cognitive and neuropsychological test batteries [5, 6]. Originally, the DSST was designed as a paper-and-pencil cognitive test presented on a single sheet of paper.

In the present study, user performance in a computerised version of the DSST test is analysed. The DSST was performed on a homogeneous group of participants,

K. Saeed and J. Dvorský (Eds.): CISIM 2020, LNCS 12133, pp. 243–254, 2020.
https://doi.org/10.1007/978-3-030-47679-3_21

composed of twenty six healthy students aged 20–24. The data analysed in the study originate from two boards of the DSST differing in their difficulty.

The literature proves that eye-tracking features might be applied in prediction of cognitive states. Benfatto et al. [7] used eye-tracking combined with machine learning in detecting psychological disorders. In [8] and in [9] eye-movement features were applied in the classification of visual tasks. Eye-tracking was applied in order to assess the workload and performance of skill acquisition [10]. Other studies examined cognitive workload using eye-tracking features among such groups as surgeons [11] or pilots [12].

In statistical and correlation analysis the parameters such as pupil dilation or pupil diameter size are the most often used to distinguish the state of cognitive workload [13, 14]. Additionally, such features as fixation rate and duration, saccade duration and amplitude or the number of blinks can be used in statistical analysis in the context of cognitive workload [15, 16].

The aim of the study is to verify whether features based on eye tracking might be used to classify cognitive workload level in the DSST test. The evaluation is based on eye-tracking features (fixations and saccades, blinks and pupil size) and test results (total number of matches and error ratio). The novelty of the paper is focused on the classification rate of cognitive workload level based on eye-tracking features.

The rest of the paper is structured as follows. The research procedure covering the computer application, equipment and experiment details is discussed in Sect. 2. Section 3 presents the methods applied in data processing, classification and statistical analysis procedures. The results are discussed in Sect. 4, whereas conclusions are presented in Sect. 5.

2 The Research Procedure

2.1 The Computer Application

The Digit Symbol Substitution Test (DSST) applied in the study was a computerised version of the DSST developed on the basis of the original paper-and-pencil cognitive test [15, 16]. The test requires a subject to match symbols to numbers according to a key located at the bottom of the screen. A symbol is assigned by clicking it on the key. A currently active letter is marked with a graphical frame. After assigning a symbol to a letter, the frame is moving to the next letter. The subject matches symbols to subsequent letters within specified time. Subsequent letters were generated randomly, with repetitions and continuously within a defined period of time.

The number of symbols and the time is defined in the application settings. In the case study two DSST parts were applied:

- 4 different symbols to assign; the test lasted 90 s.
- 9 different symbols to assign; the test lasted 180 s.

The application was developed in Java and is operated using a computer mouse. The interface of the computerised version is presented in Fig. 1.

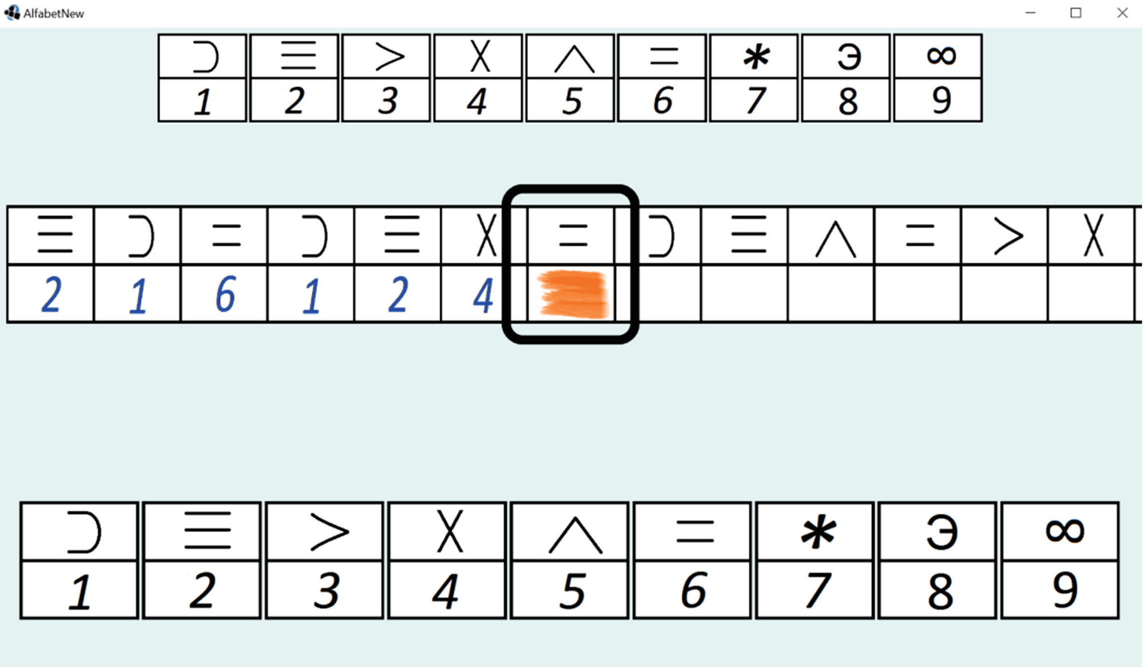


Fig. 1. The interface of the application.

2.2 Setup and Equipment

The experiment was conducted in a laboratory, in a testing room illuminated with standard fluorescent light. Eye activity was recorded using screen-based eye tracker Tobii Pro TX300 (Tobii AB, Sweden). The Tobii Pro TX300 uses video-oculography based on the dark pupil and corneal reflection method. It collects binocular gaze data with the frequency of 300 Hz.

The experiment was designed in Tobii Studio 3.2, the software compatible with the Tobii Pro TX300 eye tracker, dedicated for preparing and analysing eye-tracking experiments. Visual stimuli were presented on a separate monitor (23” TFT monitor at 60 Hz). During the experiment the participants were seated at a distance from the screen between 50 and 80 cm. The differences were insignificant for the results and they were depended on individual participant preferences (a comfortable position for working with a computer) and All participants were tested using the same software and hardware settings.

2.3 Experiment

The experiment was conducted in a dedicated laboratory with eye-tracker and computer. The 26 participants spanned the age range of 20 to 24 (mean = 20.77 years, std. dev. = 1.65). A single participant was examined for approximately 15 min. The experiment was divided into two parts, with calibration before each part.

At the beginning of each session a 9-point built-in calibration procedure was run on the eye-tracker. Then, the participants were provided with the instructions on the screen, in which they were asked to make as many matches as possible by assigning symbols to the appearing letters. The assigning was to be done by clicking a key with a particular symbol. Next, the participants completed two parts of the DSST using the computer application. Each part had a different number of symbols to assign and lasted a different amount of time. At the beginning of each part, a short trial, consisting of 9 symbols, was run to familiarise the participants with the task. After the trial, the proper test was started. Figure 2 presents the procedure of the experiment.

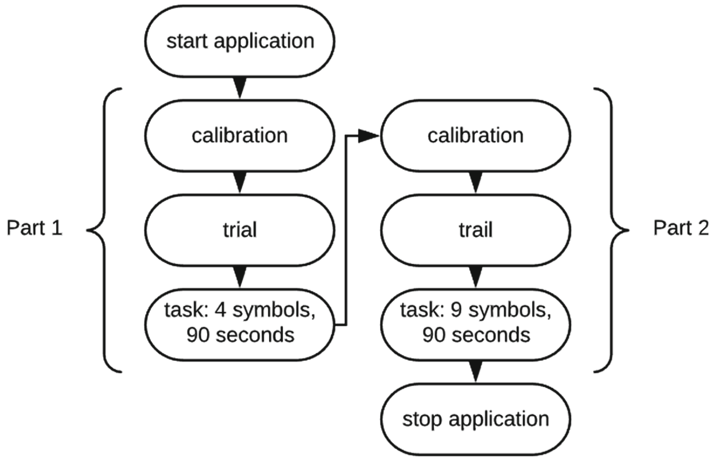


Fig. 2. The procedure of the experiment.

2.4 The Data Set

The dataset consists of 52 files generated from the eye-tracker and 52 files from the application. The total number of files generated per participant is (equal to) 4:2 files from the eye-tracker and 2 from the application. The data from each task were saved in a separate file

3 Applied Methods

3.1 Data Processing

Figure 3 presents the procedure of data processing. The procedure of data processing is divided in several main steps: data acquisition, data pre-processing, feature extraction, feature selection and the classification process.

The pre-processing step consisted of data synchronisation. Four files were generated for one participant. Since two files were generated for each part: one file from the computer application and one from the eye-tracker, the synchronisation procedure was necessary to prepare the dataset. The data from the two files were synchronised on the basis of the time stamps contained in the files. One file per part for each participant was created as the result of the synchronisation. After pre-processing, the feature selection step was performed, during which twenty features were extracted. These features are: the mean, standard deviation, maximum value, minimum value, duration of fixations and saccades and the maximum amplitude of saccades. The mean and standard deviation were calculated for the left and right pupil. The number of blinks and duration of

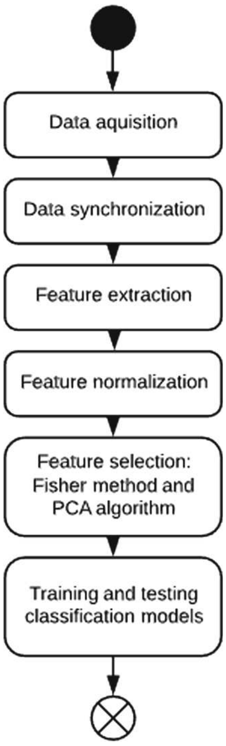

Data aquisition
Data synchronization
Feature extraction
Feature normalization
Feature selection: Fisher method and PCA algorithm
Training and testing classification models

Fig. 3. The procedure of data processing.

a blink were also extracted. Additionally to eye-tracking features, a DSST-based set of features were obtained: number of responses, number of error responses and mean time of responses. The next step in the procedure of data processing was feature normalisation, which was necessary to ensure a uniform scale of the features.

In order to reduce the input dimension number and ensure higher classification accuracy, feature selection was performed. Two methods of feature selection were applied in the analysis:

- Fisher score feature selection method [17] to select the most valuable features, and
- Principal Component Analysis (PCA) [18] to find principal components with high variance.

The following classification models were applied: The following features appeared at the top of the ranking: standard deviation of the right pupil, the mean of the left pupil, standard deviation of the saccades, the mean blink duration and the number of blinks.

After the selection feature step was completed, the final step – training and classifying was started. The following classification models were applied

- Support Vector Machine (SVM) with linear kernel
- Support Vector Machine (SVM) with polynomial (poly) kernel
- Support Vector Machine (SVM) with radial based function (rbf) kernel
- K nearest neighbours (kNN)
- Random forest

The dataset was shuffled in random order and divided into train and test datasets. The test part was 20% of the entire dataset. After the learning process, the correctness of the classifier was tested using the test dataset.

3.2 Statistical Analysis

The Kolmogrov-Smirnov (K-S test) test was performed for all 20 features to determine whether the variables have a normal distribution. In order to compare the mean values from two DSST parts, the independent-samples t-Test was used. Furthermore, the Pearson correlation coefficients between each features was calculated. The analysis was performed in the MATLAB software using the Statistical and Machine Learning Toolbox.

4 Results

4.1 Classification Results

A two class classification was conducted. Observations with a low level of cognitive workload were labelled as class 1, whereas high level observations were grouped in class 2. Two approaches are presented below: the first one is based on the feature selection method and the second resorts to the application of the PCA algorithm. The following classifiers were chosen: SVM with linear kernel, SVM with poly kernel, SVM with RBF kernel, KNN and random forest. Each learning process was repeated 200 times. The number of repetitions was established on the basis of simulation of the results presented in Fig. 4. It can be observed that 200 repetitions is enough for the partial standard deviation of the partial mean (1) of classification accuracy to reach the value of 0.01. The partial mean of classification accuracy is defined as:

$$mean(acc)_i = \frac{1}{i}\sum\nolimits_{j=1}^{i} acc_j, i \in \mathbb{N},\ i \leq n \qquad (1)$$

The partial standard deviation of $\mathrm{std}(\mathrm{mean}(\mathrm{acc}))_\mathrm{i}$ is defined as the standard deviation of first i partial means.

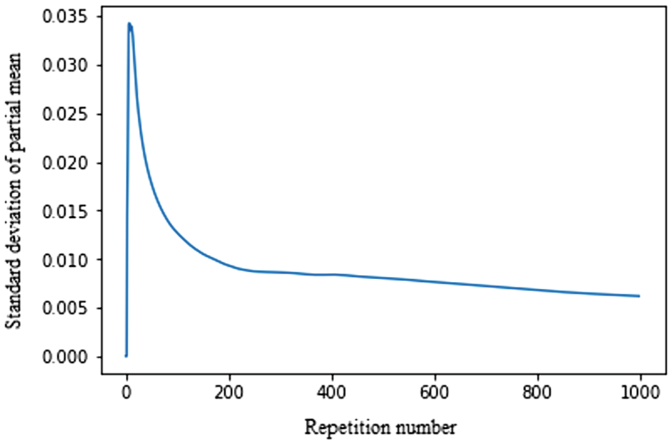


Fig. 4. Selection of repetition number – standard deviation of mean(acc).

Table 1 presents the results obtained for selected classifiers for 9 features selected by the Fisher score based feature selection method. The accuracy was calculated for each classifier. The best accuracy score was obtained for the SVM classifier with linear kernel – 0.94 and for random forest – 0.93. The worst result was obtained for the SVM classifier with poly kernel – 0.79. Tables 2 and 3 present the mean confusion matrix for the SVM with linear kernel and for random forest. Table 2 shows that on average 5.33 observations coming from the first class were classified in the proper way and only 0.285 observations from the first class were classified as second class observations. On average 5.055 observations from the second class were classified properly and 0.33 observations from the second class were classified as a first class. A similar situation occurred with the random forest confusion matrix.

Table 1. Selected classifier accuracies for 9 features selected by Fisher score based feature selection method.

Classifier	Type	Accuracy
SVM	Linear	0.94
	Poly	0.79
	Rbf	0.89
KNN		0.84
Random forest		0.93

Table 2. Confusion matrix for SVM classifier with linear kernel.

	Class 1	Class 2
Class 1	5.33	0.285
Class 2	0.33	5.055

Table 3. Confusion matrix for random forest classifier with linear kernel.

	Class 1	Class 2
Class 1	5.01	0.385
Class 2	0.375	5.23

Table 4 presents the results obtained for selected classifiers for 2 principal components. Accuracy was calculated for each classifier. Application of the PCA algorithm instead of feature selection methods ensured a high accuracy score obtained using feature selection methods and allowed to obtain even better results. The best accuracy was calculated for the SVM classifier with linear kernel. However, all the results obtained are acceptable.

Table 4. Selected classifiers accuracies for 2 principal components.

Classifier	Type	Accuracy
SVM	Linear	0.97
	Poly	0.93
	Rbf	0.93
KNN		0.95
Random forest		0.95

Figure 5 presents an example of the scatter plot for two first principal components. It can be observed that class 1 and class 2 are separable both for the train and test set observations.

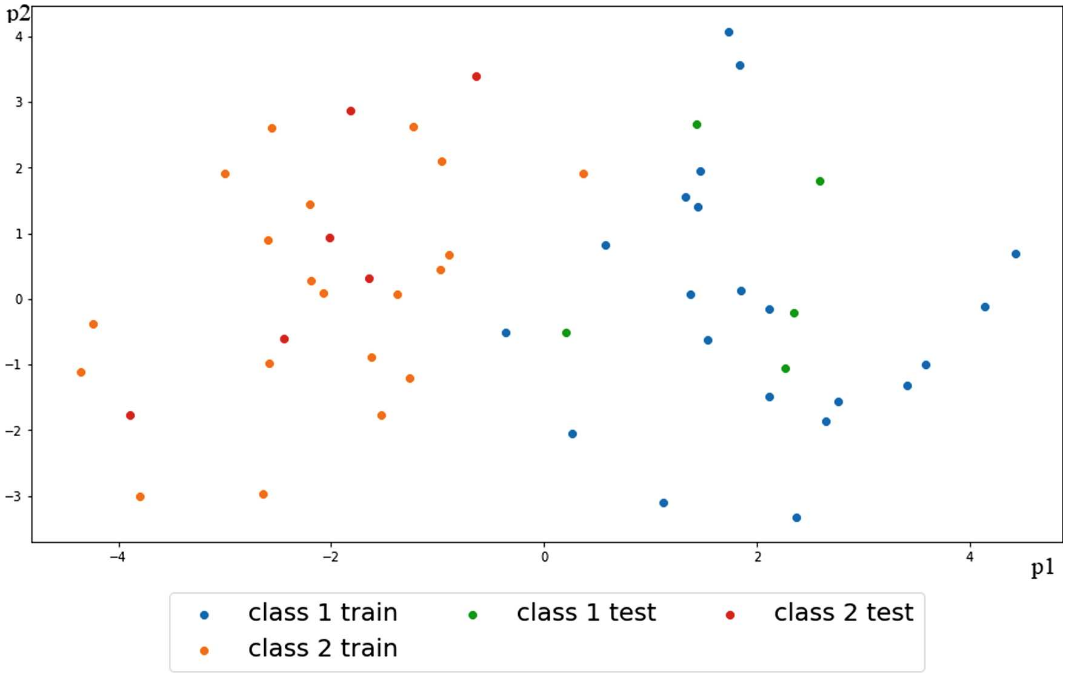


Fig. 5. Scatter plot with two first principal components.

4.2 Statistical Analysis

The K-S test found four features with non-normal distribution (Min Fix, Mi Saccade, Blinks No and Error No). The number of responses was not included in statistical analysis. The independent-samples t-Test revealed statistically significant differences for some features. Table 5 presents features revealed with the t-Test for p-value 0.05.

Table 5. The results of independent-samples t-Test

Features	P_{value}	Features	P_{value}
Mean Response Time	<0.001	Std Saccade	0.034
Fix No	<0.001	Max Saccade	<0.001
Max Fix	0.005	Mean Saccade Amplitude	<0.001
Std Fix	0.028	Std Pupil Left	0.006
Saccade No	<0.001	Std Pupil Right	0.015

Tables 6 and 7 present the statistically significant (p-value 0.05) correlation coefficients for first and second part of the DSST examination. In the case of the first DSST part, 11 pairs of correlated features were observed. The strongest correlation was observed for the Mean Pupil Right – Mean Pupil Left pair. The Blinks No – Saccade No pair also presents a high value of correlation coefficient.

The second part of the examinations revealed 11 pairs of correlated features. As in the first part of the DSST examination, the highest correlation has been found for the Mean Pupil Left – Mean Pupil Right pair. The most frequent feature to appear in the first and second examination is Blinks No, which is correlated with Mean Saccade Amplitude, Saccade No, Std Saccade No and Std Pupil Left.

Table 6. The values of correlation coefficient for first part of examination

Features	Correlation coefficient	Features	Correlation coefficient
Mean Fix Duration-Saccade No	−0.5814	Mean Saccade Duration-Std Pupil Left	−0.5002
Max Saccade-Std Fix	−0.5163	Blinks No-Mean Saccade Amplitude	−0.6109
Max Saccade-Max Fix	−0.3989	Blinks No-Saccade No	0.7049
Mean Pupil Right-Mean Pupil Left	0.9127	Blinks No-Std Saccade	0.7144
Mean Pupil Right-Std Pupil Left	0.5637	Blinks No- Std Pupil Right	0.4728
Std Pupil Left-Std Pupil Right	0.7538		

Table 7. The values of correlation coefficient for second part of examination

Features	Correlation coefficient	Features	Correlation coefficient
Mean Fix Duration-Saccade No	−0.5079	Mean Pupil Right-Std Pupil Left	0.4971
Saccade No-Max Fix	−0.5717	Std Pupil Left-Std Pupil Right	0.5776
Max Saccade-Std Pupil Left	0.4145	Std Pupil Right-Saccade No	0.4355
Max Saccade-Std Pupil Right	0.4073	Blinks No-Std Saccade	0.5874
Mean Blinks Duration-Max Fix	0.4000	Blinks No-Mean Saccade Amplitude	0.4933
Mean Pupil Right-Mean Pupil Left	0.9014		

5 Discussion and Conclusions

The aim of the paper was to verify whether eye tracking-based features might be used to classify cognitive workload level. Mental fatigue was measured during two sessions of the DSST test run in different conditions.

The binary classification was performed with different machine learning models based on such algorithms as the SVM with kernels: linear, poly and radial basis function, KNN and random forest. The evaluation was based on eye-tracking features (mean, standard deviation, maximum and minimum value, duration of fixations and saccades, maximum amplitude of saccades, the number and duration of blinks, and pupil size) and test results (the number of responses and error responses and the mean time of responses). The Fisher score feature selection method was applied to select nine of the most informative features used to build models. The learning process for each model was repeated 200 times.

The results show that the highest accuracy was achieved for the linear SVM model (94%), although the random forest algorithm (93%) occurred to be efficient as well. Confusion matrices for these models, where type I and type II errors are at relatively low levels, proved the stability of the models. Data analysis with the PCA algorithm for the first two principal components showed linear separability of classes, which corresponds to the fact that the linear model occurred to be the most efficient. The worst classification results was reached for the SVM with polynomial kernel.

Statistical analysis revealed the most significant features, which are: mean response time, standard deviation of saccades, standard deviation of fixation, fixation number, maximum saccade, maximum fixation, mean saccade amplitude, standard deviation of the right and left pupil. Most of these features were found by the Fisher score feature selection method. Statistical analysis did not reveal very strong significant correlations between features.

References

1. Gevins, A., Smith, M.E., McEvoy, L., Yu, D.: High-resolution EEG mapping of cortical activation related to working memory: effects of task difficulty, type of processing, and practice. Cereb. Cortex **7**, 374–385 (1997)
2. Boake, C.: From the Binet-Simon to the Wechsler-Bellevue: tracing the history of intelligence testing. J. Clin. Exp. Neuropsychol. **24**, 383–405 (2002)
3. Wechsler, D.: The Measurement of Adult Intelligence. The Williams & Wilkins Company, Baltimore (1939)
4. Jaeger, J.: Digit symbol substitution test: the case for sensitivity over specificity in neuropsychological testing. J. Clin. Psychopharmacol. **38**(5), 513 (2018)
5. Sicard, V., Moore, R.D., Ellemberg, D.: Sensitivity of the Cogstate Test Battery for detecting prolonged cognitive alterations stemming from sport-related concussions. Clin. J. Sport Med. **29**(1), 62–68 (2017)
6. Cook, N.A., et al.: A pilot evaluation of a computer-based psychometric test battery designed to detect impairment in patients with cirrhosis. Int. J. Gen. Med. **10**, 281–289 (2017)
7. Benfatto, M.N., Seimyr, G.Ö., Ygge, J., Pansell, T., Rydberg, A., Jacobson, C.: Screening for dyslexia using eye tracking during reading. PLoS One **11**(12) (2016)
8. Coco, M.I., Keller, F.: Classification of visual and linguistic tasks using eye-movement features. J. Vis. **14**(3), 11 (2014)
9. Henderson, J.M., Shinkareva, S.V., Wang, J., Luke, S.G., Olejarczyk, J.: Predicting cognitive state from eye movements. PLoS ONE **8**(5), 1–6 (2013)
10. Mark, J., et al.: Eye tracking-based workload and performance assessment for skill acquisition. In: Ayaz, H. (ed.) AHFE 2019. AISC, vol. 953, pp. 129–141. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-20473-0_14
11. Ortega-Morán, J.F., Pagador, J.B., Luis-del-Campo, V., Gómez-Blanco, J.C., Sánchez-Margallo, F.M.: Using eye tracking to analyze surgeons' cognitive workload during an advanced laparoscopic procedure. In: Henriques, J., Neves, N., de Carvalho, P. (eds.) MEDICON 2019. IP, vol. 76, pp. 3–12. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-31635-8_1
12. Van Acker, B.B., et al.: Mobile pupillometry in manual assembly: a pilot study exploring the wearability and external validity of a renowned mental workload lab measure. Int. J. Ind. Ergon. **75** (2020). https://doi.org/10.1016/j.ergon.2019.102891
13. Marshall, S.P., Pleydell-Pearce, C.W., Dickson, B.T.: Integrating psychophysiological measures of cognitive workload and eye movements to detect strategy shifts. In: Proceedings of the 36th Annual Hawaii International Conference on System Sciences, Big Island, HI, USA, p. 6 (2003)
14. Marshall, S.P.: The index of cognitive activity: measuring cognitive workload. In: Proceedings of the IEEE 7th Conference on Human Factors and Power Plants, Scottsdale, AZ, USA, p. 7 (2002)
15. Chen, S., Epps, J., Ruiz, N., Chen, F.: Eye activity as a measure of human mental effort in HCI. In: Proceedings of the 16th International Conference on Intelligent User Interfaces, Palo Alto, CA, USA, pp. 315–318 (2011)
16. Tokuda, S., Obinata, G., Palmer, E., Chaparro, A.: Estimation of mental workload using saccadic eye movements in a free-viewing task. In: 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 4523–4529. IEEE Engineering in Medicine and Biology Society (2011)

17. Gu, Q., Li, Z., Han, J.: Generalized fisher score for feature selection. arXiv preprint arXiv:1202.3725 (2012)
18. Pechenizkiy, M., Tsymbal, A., Puuronen, S.: PCA-based feature transformation for classification: issues in medical diagnostics. In: Proceedings of the 17th IEEE Symposium on Computer-Based Medical Systems, pp. 535–540. IEEE (2004)

6.2 Interpretable Machine Learning Models for Three-Way Classification of Cognitive Workload Levels for Eye-Tracking Features

1. Details

This article was written by Monika Kaczorowska, Małgorzata Plechawska-Wójcik and Mikhail Tokovarov and published in Brain sciences in 2021, vol. 11, nr 2. The article is worth 100 points according to the list of the Polish Ministry of Science and Higher Education.

2. Abstract

The paper is focused on the assessment of cognitive workload levels using selected machine learning models. In this study, eye-tracking data were gathered from 29 healthy volunteers during the examination with three versions of the computerized version of the digit symbol substitution test (DSST). Understanding cognitive workload is of great importance in analyzing human mental fatigue and the performance of intellectual tasks. It is also essential in the context of explanation of the cognitive context of the brain. Eight three-class classification machine learning models were constructed and analyzed. Furthermore, a technique of interpretable machine learning model was applied to obtain the measures of feature importance and its contribution to the brain's cognitive functions. The measures allowed the improvement of the quality of classification, simultaneously lowering the number of applied features to six or eight, depending on the model. Moreover, the applied method of explainable machine learning provided valuable insights into understanding the process accompanying various levels of cognitive workload. The main classification performance metrics, such as F1, recall, precision, accuracy, and the area under the Receiver operating characteristic curve (ROC AUC) were used in order to assess the quality of the classification quantitatively. The best result obtained on the complete feature set was as high as 0.95 (F1); however, feature importance interpretation allowed to increase the result up to 0.97 with only seven of the 20 features applied.

brain sciences

MDPI

Article

Interpretable Machine Learning Models for Three-Way Classification of Cognitive Workload Levels for Eye-Tracking Features

Monika Kaczorowska , Małgorzata Plechawska-Wójcik * and Mikhail Tokovarov

Department of Computer Science, Lublin University of Technology, 20-618 Lublin, Poland; m.kaczorowska@pollub.pl (M.K.); m.tokovarov@pollub.pl (M.T.)

* Correspondence: m.plechawska@pollub.pl

Abstract: The paper is focussed on the assessment of cognitive workload level using selected machine learning models. In the study, eye-tracking data were gathered from 29 healthy volunteers during examination with three versions of the computerised version of the digit symbol substitution test (DSST). Understanding cognitive workload is of great importance in analysing human mental fatigue and the performance of intellectual tasks. It is also essential in the context of explanation of the brain cognitive process. Eight three-class classification machine learning models were constructed and analysed. Furthermore, the technique of interpretable machine learning model was applied to obtain the measures of feature importance and its contribution to the brain cognitive functions. The measures allowed improving the quality of classification, simultaneously lowering the number of applied features to six or eight, depending on the model. Moreover, the applied method of explainable machine learning provided valuable insights into understanding the process accompanying various levels of cognitive workload. The main classification performance metrics, such as F1, recall, precision, accuracy, and the area under the Receiver operating characteristic curve (ROC AUC) were used in order to assess the quality of classification quantitatively. The best result obtained on the complete feature set was as high as 0.95 (F1); however, feature importance interpretation allowed increasing the result up to 0.97 with only seven of 20 features applied.

Keywords: cognitive workload; mutliclass classification; explainable machine learning; eyetracking signal

check for updates

Citation: Kaczorowska, M.; Plechawska-Wójcik, M.; Tokovarov, M. Interpretable Machine Learning Models for Three-Way Classification of Cognitive Workload Levels for Eye-Tracking Features. *Brain Sci.* **2021**, *11*, 210. https://doi.org/10.3390/brainsci11020210

Academic Editor: Kip Ludwig

Received: 17 December 2020

Accepted: 3 February 2021

Published: 9 February 2021

1. Introduction

Understanding cognitive workload as a mental effort needed to perform a task [1] is important in human mental fatigue analysis. The diverse complexity of mental tasks requires different levels of concentration. Their understanding and categorising might be useful in the process of modelling information processing capabilities. The level of mental fatigue and its influence on the brain cognitive capability is the subject of numerous research articles [2,3]. Mental fatigue might lead to a decrease of brain cognitive system performance in terms of perception, attention, analysing, and planning [4,5]. What is more, mental fatigue might affect reaction times, target-detection failure, and other objective declines [6].

The literature review conducted shows that the most widely used method of assessment of cognitive workload level in the past employed subjective measures, such as NASA Task Load Index (NASA-TLX) [7,8]. However, the psycho-physiological state might be assessed by objective methods based on bio-signals, such as the eye-tracking technique [9], Galvanic Skin Response (GSR) [10,11], electroencephalogram (EEG), pupillometry, or electrocardiogram (ECG) [12].

Eye-tracking data turn out to be useful in analysing cognitive workload [13]. Benfatto et al. [14] apply eye-tracking features to detect psychological disorders. Workload and performance of skill acquisition based on eye-tracking data are presented in [15]. Eye-movement based classification of visual and linguistic tasks is discussed in [16]. There are also numerous studies presenting cognitive workload classification with a combination of eye-tracking and other bio-signals, such as EEG [17], although the literature does not present numerous cognitive workload classification studies based only on eye-tracking data.

Most cognitive workload classification studies presented in the literature are binary approaches. Among them, the Support Vector Machines (SVM) classifier is one of the most popular [18–20]. The above studies, based on different bio-signal data, report its accuracy at the level of even 94–97%. Other popular methods are Linear Discriminant Analysis (LDA) [21], k-Nearest Neighbours (kNN) [22], or Multilayer Perceptron (MLP) [23]. Multiclass problem studies can also be found. Authors applied such methods as SVM ([24] 71%), linear regression ([25] accuracy 82%), or neural networks ([26] 74%, [27] 83%).

Most cognitive workload classification results were achieved in a classical subject-specific approach [27], for which researchers report higher classification performance [20,28]. However, developing a subject-independent classifier allows one to distinguish between cognitive workload levels regardless of external and internal conditions such as the age, time of day, or habits of an examined person. Nevertheless, the literature presents only a few publications with subject-independent approaches [26–28]. In [29], the authors conducted a subject-independent and subject-dependent classification based on EEG signals from 14 participants. Thodoroff et al. created a classifier model based on a subject-independent approach [30] using a dataset containing 23 patients.

The aim of the present study includes the following points:

- Performing a multiclass subject-independent classification of cognitive workload levels,
- Examine both classification on the complete feature set and with the application of interpretable machine learning models for feature selection,
- Carrying out a deeper analysis of the features related to the classification of particular levels of cognitive workload.

The dataset used in the study is eye-tracking and user performance data gathered from 29 participants while solving a computerised version of the digit symbol substitution test (DSST).

The digit symbol substitution test (DSST) [31] is a cognitive tool introduced as a paper-and-pencil test originally applied in order to understand human associative learning. Currently, this test is commonly applied in clinical neuropsychology to measure cognitive dysfunction and is often present in cognitive and neuropsychological test batteries [32,33]. The DSST allows one to check the patient's processing speed, memory, and executive functioning [34].

The rest of the paper is structured as follows. The review of the literature is presented in Section 2, while the research procedures covering the computer application, equipment, and experiment details are discussed in Section 3. Section 4 presents the methods applied in data processing, classification, and statistical analysis procedures. Results are discussed in Section 5. Section 6 contains an analysis and discussion of the extracted feature importance measures. Section 7 concludes the paper.

2. Related Work

Table 1 shows a review of the literature where the numbers of participants and cognitive workload levels are presented. The approach column indicates a subject-dependent (sd) or subject-independent (si) approach of classification. For the scientific articles, the results of classification are presented as well. In [35], the authors wrote about labelling cognitive workload data using three methods: difficulty split by expert, the Rash model, and the stress–strain model. The most common method is based on task difficulty conditions named difficulty split labelling, while the Rash and the stress–strain models allow adjusting the cognitive workload level to each participant separately. The authors conducted the examination on 34 participants, obtaining data to label and classify the level of cognitive workload [35]. Lobo et al. published a study on the classifying levels of cognitive workload based on eye-tracker and EEG data [17]. They conducted a three-class classification where they created three levels of cognitive workload and applied a subject-independent approach based on 21 observations. Both the experts and novices took part in the experiment [36], having a low level of cognitive workload while the novices had a higher one [36]. The authors conducted a two-class subject-independent classification based on data from 14 participants. In [37], the authors carried out the experiment asking 35 participants to play two games with different levels of difficulty.

Almogbel et al. published a study examining the levels of cognitive workload in the process of playing a computer game [38]. The data were collected from one participant, and a classification of low and high level of cognitive workload was obtained. The same authors published another paper where they tried to classify three and six levels of cognitive workload on the basis of data from one participant in the process of playing a computer game [39]. In [38,39], the authors applied a subject-dependent approach to create a classification model. Study [40] attempted to classify low and high levels of cognitive workload defined on the basis of data gathered from eight participants. In [41], a subject-independent classifier was made on the basis of data from 12 participants. The experiment was conducted using arithmetical tasks defining seven levels of cognitive workload, where the first level was the easiest and contained one and two-digit numbers, and the seventh level was related to arithmetical tasks on three-digit numbers with three carries. The authors attempted to classify the three and two cognitive workload levels using a subject-independent and dependent approach based on the pupil data from 25 participants [42]. In [43], two approaches were applied to create a classification model as well. The authors considered the prediction of a driver’s cognitive workload: good or poor driving performance state while driving based on EEG data from 37 participants. In [44], the authors applied eye tracking in virtual reality (VR) and augmented reality (AR) technology to classify the cognitive load of drivers under critical situations. Two-class subject-independent classification was conducted using several types of classifiers: SVM, decision tree, random forest, and k-Nearest Neighbours and used five metrics: accuracy, precision, recall, and F1-score. Data were gathered from 16 participants and two levels of cognitive load were defined: low and high.

Fridman and colleagues [45] conducted the experiment based on making videos of real-time cognitive load in various contexts which were corresponding to cognitive load level. A total of 92 participants took part in the experiment, and three-level classification was applied using convolutional deep neural network 3D and Hidden Markov Model. In [45], a subject-independent approach was used to create the classification model based on eye images extracted from videos. In [46], the authors noticed that most of the previous research was based on data not related to older adults. They mentioned that the changes in eye-tracking could appear in older adults, so they conducted an experiment where older adults were asked to watch the video clips. A two-level classification model was created on the basis of the data gathered from 12 participants. The model is able to classify the fatigue and non-fatigue state independently from the participant’s age. In [47], the authors presented an interesting approach to cognitive workload estimation using EEG signals with the application of a deep convolutional neural network with residual connections and a gated recurrent unit (GRU). A high accuracy of subject-independent classification was reported for an approach with four workload levels. In [48], a model capable of distinguishing between two cognitive workload levels was developed. The authors proposed a novel approach consisting of two steps that included the initial training of a set of participant-specific classifiers and then combining the trained classifiers to address the problem of subject-independent cognitive workload estimation. In [49], Custom Domain Adaptation (CDA) was used to develop a highly efficient classifier, which was trained with the same dataset as it was applied in [47].

Table 1. A review of the literature.

Literature	Number of Participants	Number of Cognitive Workload Levels	Classifier Model	Result	Approach (sd/si)
		difficulty split		0.51 (AUC)	
[35]	34	3 Rash model	Random forest	0.81 (AUC)	si
		stress–strain model		0.67 (AUC)	
[17]	21	3, difficulty split	kNN,	0.332 (F score)	si
[36]	14	2, difficulty split	LDA	0.91 (ACC)	si
[37]	35	2, difficulty split	SVM	0.52 (ACC)	si
[38]	1 (24 recordings)	2, difficulty split	Convolutional deep neural networks	0.95 (ACC)	sd
[39]	1 (24 h of recordings)	2, difficulty split	Convolutional deep neural network	0.934 (ACC)	sd
		3, difficulty split		0.976 (ACC)	
		6, difficulty split		0.945 (ACC)	
[40]	8	2	Deep neural network	0.868 (ACC)	si
[41]	12	7, difficulty split	Artificial Neural Network	0.4–0.98 (ACC)	si
[42]	25	2, difficulty split	Extra Trees	0.768 (ACC)	si
				0.824 (ACC)	sd
		3, difficulty split		0.467 (ACC)	si
				0.637 (ACC)	sd
[43]	37	2, difficulty split	Convolutional deep neural network	0.767 (ACC)	si
				0.861 (AUC)	sd
[44]	16	2, difficulty split	SVM	0.81 (AUC)	si
[45]	92	3, difficulty split	Convolutional deep neural network 3D	0.86 (AUC)	si
[46]	12	2, difficulty split	SVM	0.92 (AUC)	si
[47]	13	4, difficulty split	Deep neural network	0.907 (ACC)	si
				0.896 (F score)	
[48]	47	2, difficulty split	Forest of Extremely Randomised Trees	0.72 (ACC)	si
[49]	13	4, difficulty split	Deep neural network	0.98 (AUC)	si

3. The Research Procedure

3.1. The Computer Application

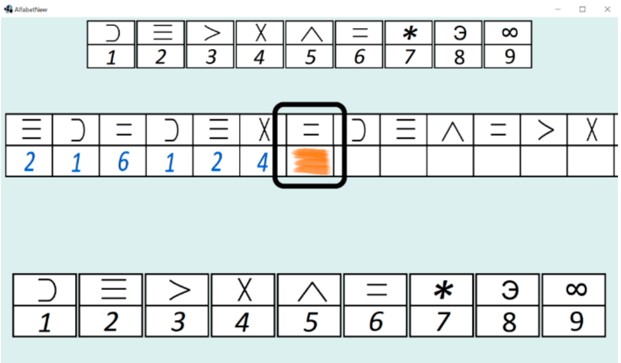
In the present study, the computerised version of the DSST test [50,51] was applied. A subject was asked to match symbols to numbers according to a key located at the bottom of the screen. To assign a particular symbol, a subject had to click on the key corresponding to this symbol. A graphical frame (Figure 1) marked the currently active letter, and after assigning a symbol, the frame moved to the next letter. The subject matched symbols to subsequent letters within the specified time. Subsequent letters were generated randomly and with repetition, and they appeared continuously within the defined time.

The number of symbols and the time is defined in the application settings. In the case study, three DSST parts were applied:

- Part 1—low-level cognitive workload: four different symbols to choose from, 90 s test length;
- Part 2—medium-level cognitive workload: nine different symbols to choose from, 90 s test length;
- Part 3—hard-level cognitive workload: nine different symbols to choose from, 180 s test length.

These three parts correspond to the classes of cognitive workload, defined in Section 3.3. Three levels of cognitive workload were empirically defined and separated on the basis of a preliminary pilotage study performed on a small group of two participants with a profile consistent with the characteristics of the study participants (they did not participate in the target study). The number of symbols and the duration of the test were set in an interview carried out after this preliminary examination.

The Java 8.0 programming language was the main tool used for developing the application; it was designed to be operated by a computer mouse. The interface of the computerised version is presented in Figure 1. The legend containing the characters used and the digits assigned to them is shown at the top in the form of the table. The user can select a specific character by clicking the proper cell in the table at the bottom of the window. The central part contains the currently active task (a user is supposed to click the character with digit 6). The left part of the central table presents the history of the tasks, and the right part shows the upcoming tasks.



Ↄ	≡	>	X	∧	=	*	Э	∞
1	2	3	4	5	6	7	8	9

≡	Ↄ	=	Ↄ	≡	X	=	Ↄ	≡	∧	=	>	X
2	1	6	1	2	4							

Ↄ	≡	>	X	∧	=	*	Э	∞
1	2	3	4	5	6	7	8	9

Figure 1. The interface of the application. The highlighted area presents the current symbol to be matched.

3.2. Setup and Equipment

The described experiment was carried out in a laboratory with standard fluorescent light. In order to ensure stable, equal conditions for all participants, the outside light was blocked. The activity of eyes was registered by a Tobii Pro TX300 screen-based eye tracker (Tobii AB, Stockholm, Sweden) utilising near-infrared technology [52]. The video-oculography method in the said device is based on corneal reflection as well as the dark pupil method (VOG). It collects the data related to binocular gaze with the frequency of 300 Hz for studies of saccades, fixations, and blinks. The sampling rate variability is less than 0.3%. The gaze precision (binocular) is 0.07°, and the gaze accuracy is 0.4°. The eye tracker is built into a monitor (23″ TFT monitor at 60 Hz) connected to the computer (laptop computer Asus G750JX with 8 GB of RAM and processor Intel Core i7–4700HQ) on which the experiment was carried out. Tracking is proceeded for each eye separately.

The experiment was designed in Tobii Studio 3.2, which is the software compatible with the Tobii Pro TX300 eye tracker, and it is dedicated to preparing and analysing eye-tracking tests. Visual stimuli were presented on a separate monitor (23″ TFT monitor with 60 Hz of refresh rate). During the experiment, the participants occupied a seated position in such a way that the screen was from 50 to 80 cm away from the participant depending on the fact of which position was convenient for the participant to work with the computer. The experiment was preceded by the nine-point calibration procedure realised for each participant, so the distance from the monitor within the mentioned range did not affect the results of the experiment. The calibration procedure is performed separately for each eye. Identical hardware and software settings were used for examining all the participants.

The Tobii Studio was applied to export eye activities gathered during the experiment. The signal lost was used to check the tracking ratio to ensure the proper quality of data. Eye activities exported from the experiments were related to several measures:

- Fixations [53], originally defined as the period of uptake of visual information, when a participant holds eyes relatively stable in a particular position. Single fixation occurs between two consecutive saccades. All visual input occurs [54] during fixations.
- Saccades [53] are defined as the rapid eye movement occurring between fixations. During saccades, the eye gaze is moved from one point to another to bring the part of visual information onto the most sensitive part of the retina in order to retrieve information [54].
- Blinks derived as zero data embedded in two saccadic events.
- The blink was identified as zero data is embedded in two saccadic events [46,55,56].
- Pupillary response understood as pupil size. The Tobii Studio applies the dark pupil eye-tracking method.

Fixations and saccades are detected in the Tobii Studio with the Velocity–Threshold Identification (I-VT) fixation classification algorithm [57], which classifies the eye movements using directional velocity shifts of the eye. The velocity threshold parameter was set to 30°/s [58]. This is the recommended value sufficient for recording with the average level of noise. The research on the I-VT velocity threshold parameter for the Tobii TX300 eye-tracker show that it provides stability when the I-VT threshold is between 20°/s and 40°/s [52]. The amplitude of saccades and blinks were determined directly from the eye tracker via 3D eye position and screen gaze points.

3.3. Experiment

The experiment was applied on a group of 29 participants aged 20 to 24 (mean = 20.61 years, std. dev. = 1.54). The tested group was homogeneous. Healthy participants (23 males, six females) were recruited among second- and third-year students of the BSc degree in computer science. The participants had normal/corrected to normal vision. The acceptable level of data activity recorded by the eye-tracker was set to 90%. Originally, 30 participants were invited to participate in the study; however, one participant was discarded due to poor data eye-tracking quality caused by excessive body and head movement (the data activity level recorded by the eye-tracker for this participant was 70%).

The research was approved by the Committee of the Lublin University of Technology (Approval ID 2/2018 from 19 October 2018). The participation was voluntary, and all participants received information about the study.

A single participant was examined for approximately 15 min. The experiment was divided into three parts. Each part was preceded by a calibration phase. The calibration phase was run on the eye-tracker as a built-in procedure. The next step contained the instructions demonstrated to the participants on the screen. The participants were asked to create as many matches as possible by assigning symbols to the appearing letters. The assigning was done by clicking on a particular symbol on the key located at the bottom of the screen (see Figure 1). Afterwards, the participants went through the three parts of the DSST using the computer application. Every experiment stage contained a different number of symbols to assign and lasted a different period of time. Each stage started from a short trial including nine symbols, and the participants were supposed to familiarise with the task by performing the trial. After finishing the initial part, a participant could start the main part of the test stage. Figure 2 presents the procedure of the experiment.

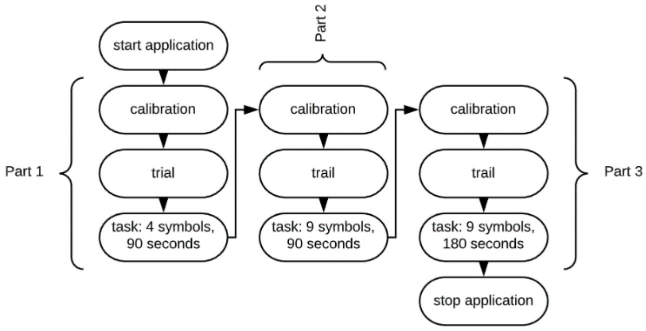


Figure 2. The procedure of the experiment.

3.4. Dataset

The dataset consists of files generated from the eye-tracker and files generated from the application. The files generated from the eye-tracker consist of the time stamps and the data that are related to such eye activity as fixations, saccades, blinks, etc. The files generated by application include the time stamps and the data that are related to DSST test results such as number of errors or number of responses. The total number of generated files per participant equals six, i.e., three files from the eye-tracker and three files from the application. The data from each task were saved in a separate file.

Eye activity-related and DSST test results-related data exported from the eye tracker covering 20 features are presented below:

- Fixation-related features: fixation number (total number of fixations), mean duration of fixation, standard deviation of fixation duration, maximum fixation duration, minimum fixation duration.
- Saccade-related features: saccade number (total number of saccades), mean duration of saccades, mean amplitude of saccades, standard deviation of saccades amplitude, maximum saccade amplitude, minimum saccade amplitude.
- Blink-related features: blink number (total number of blinks), mean of blink duration.
- Pupillary response features: mean of left pupil diameter, mean of right pupil diameter, standard deviation of left pupil diameter, standard deviation of right pupil diameter.
- DSST test results-related features: number of errors (total number of errors), mean response time, response number (total number of responses).

The listed eye-activity related features were chosen as the most informative ones available in the eye tracker software. Features related to fixations and saccades are the most common eye-tracking features analysed in the literature [54] in such areas as psychology and neuroscience, including behavioural patterns, mental fatigue, and disorders analysis [46,57]. Although in the literature the most common approaches consider only the main fixation- and saccade-related features such as the total number of saccades, mean duration of saccades, total number of blinks, and mean of blink duration, we decided to consider also an additional set of features including standard deviation of saccades amplitude, standard deviation of fixation duration, maximum and minimum saccade amplitude, and fixation duration. Even though these features are not widely applied in cognitive workload research, they were included in the analysis to check its possible usefulness in the process of classification. Standard deviation, mean and skewness of fixation duration, and saccade amplitude have been previously applied in the task of classifying mental states [59]. Standard deviations of the saccade parameters were also used in the analysis of eye-movement relation to the age of the participants [60]. Distributions of fixation durations and of saccade amplitudes were analysed also in the context of classifying eye fixations [61]. Motivation to analyse maximum and minimum values (included in the analysis after correction of outliers) as well as standard deviation of fixation duration and saccade amplitude (as a measure of variance or dispersion) was related to potential statistical differences between particular cognitive workload levels.

Blink-related features as well as pupillary response features were proved to indicate the dynamics of the cognitive process [1,62,63]. Pupil diameter-related features were measured separately for each eye and needed additional preprocessing steps. The pupil size was reported in millimeters, and it was estimated based on the magnification effect achieved by both the spherical cornea and the distance to the eye [46]. Linear interpolation was applied in order to reduce the impact of blinking or artifacts [64]. A sudden pupil size change of 0.1 mm, within a 3 ms time span, was marked as an artifact [64,65]. Missing data, especially data related to blinks, were ignored and were not included in the further processing. A subtractive baseline correction (corrected pupil size = pupil size − baseline) [66] has been applied to the pupil size data. The baseline has been estimated based on 100 ms fragment recorded before the main experimental procedure (during the welcome page displayed). The DSST test results-related features were included to complete the analysis process.

4. Methods Applied

4.1. Data Processing

Eye activity data and DSST test results were analysed off-line using custom programs written in MATLAB 2020a and Python 3.6. The procedure of data processing was divided into 6 steps:

- Data acquisition
- Data synchronisation
- Feature extraction
- Feature normalisation
- Feature selection
- Training and testing classification models.

Six files were generated per participant, three for both parts: one file from the computer application and one from the eye-tracker. The essential part of preprocessing was the procedure of synchronisation, which was necessary to prepare the dataset. It allowed combining the files from the eye-tracker and application and obtaining one file per each part of the experiment. Every participant provided three observations: the first one corresponded to a low level of cognitive workload (file 1), the second one was labelled as moderate (file 2), and the last one had the signal associated with a high level of cognitive workload (file 3). The dataset included 87 observations in total (3 observations per participant). Every observation consisted of independent features and the class label (low, moderate, or high). The values of independent features were obtained by means of a feature extraction step, leading to the twenty features listed in Section 3.4. The feature normalisation step was the next stage in the procedure of data processing; it was necessary to ensure a uniform scale of the features.

4.2. Feature Selection

In order to obtain the values of feature importance measure, the logistic regression model was used. Due to its properties, logistic regression is commonly used in the applications where deep understanding of the reasons behind the decision taken by the model is necessary [67]. The model of logistic regression can be expressed by the formulae below. First, the weighted sum of the feature values is computed as shown below:

$$S = \sum_{i=1}^{n} w_i \cdot a_i + w_0. \tag{1}$$

The result is calculated with the use of the following logistic function:

$$p = \frac{1}{1 + e^{-s}} \tag{2}$$

where

- p is the probability of the fact that the classified sample belongs to the positive class,
- i is the order number of the feature,
- w_i, $(i \in [1, n])$ is the weight of the i-th feature; the lower the absolute value, the less important the feature, and conversely, higher absolute values of the weights correspond to the features producing great influence on the model's decisions,
- w_0 is the bias,

a_i, $(i \in [1, n])$ is the value of i-th feature.The values of weights $\{w_i\}$, representing feature importance, are obtained in the process of model training. Hence, training a logistic regression model on the complete feature set is a starting point for feature selection. Figure 3 presents the process of feature selection and appliance of its results in classification. The logistic regression models applied for feature selection include also regularisation elements; after initial consideration, elastic net, being the linear combination of L1 and L2 regularisation, was applied. The approach allowed mitigating possible correlations between features, so that in a pair of correlated features, one obtained a higher importance weight and the other was assigned a notably lower weight.

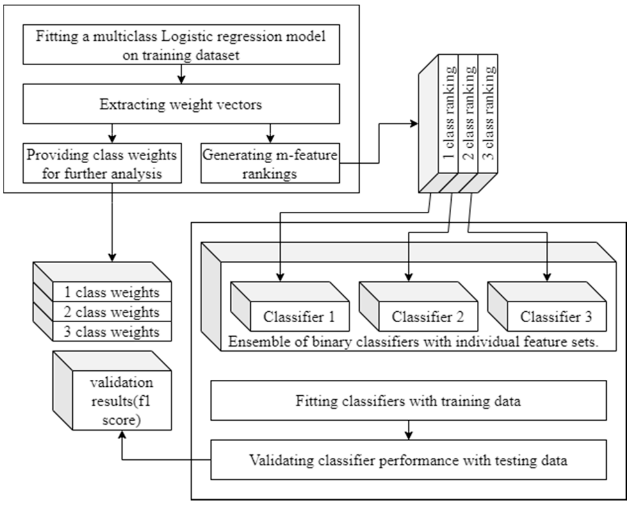

Fitting a multiclass Logistic regression model on training dataset
Extracting weight vectors
Providing class weights for further analysis
Generating m-feature rankings
1 class ranking
2 class ranking
3 class ranking
1 class weights
2 class weights
3 class weights
validation results(f1 score)
Classifier 1
Classifier 2
Classifier 3
Ensemble of binary classifiers with individual feature sets.
Fitting classifiers with training data
Validating classifier performance with testing data

Figure 3. Feature selection experiment flowchart.

4.3. Classification

The aim of the classification process was to assign observations into one of the three classes:

- Class 1—observations with low level of cognitive workload
- Class 2—observations with moderate level of cognitive workload
- Class 3—observations with high level of cognitive workload

Classes correspond to three cognitive workload levels applied in the DSST application.

Various classification methods, such as SVM, kNN, Random Forest, Multilayer Perceptron (MLP), and Logistic Regression were applied in order to produce initial results, showing the general perspectives of eye-tracking data as the source of information for cognitive workload assessment.

The classifier models mentioned had the following hyperparameters:

- kNN—nearest neighbour number: 5
- Random Forest—tree number: 100
- MLP—two hidden layers: 32 and 16 neurons; optimiser: Adam; learning rate: 0.0001; activation function: relu (rectifier linear unit)

Afterwards, logistic regression was used for extracting feature weights representing their importance in the process of distinguishing between particular levels of cognitive workload. The influence of selected features was tested for the classifiers mentioned before.

The mentioned classification models were chosen due to the following reasons: lower number of parameters (compared to deep learning models), lower tendency to overfit on small datasets (compared to deep learning models), low computational cost allowing running multiple experiments in a reasonable amount of time. Moreover, selection of algorithm was performed considering the other studies in the field of cognitive workload classification.

The dataset was shuffled in random order and divided into train and test datasets. Data from every participant could be used only in one dataset: train or test, in order to ensure a truly subject-independent approach, so that the signals of the test dataset persons would be completely unknown for the model. Approximately 20% of the input dataset, which corresponds to 6 participants, was used for testing.

A range of classical machine learning models were used. The applied approach included the initial state, where the classifiers were tested on the complete feature set and afterwards the main analysis where the feature selection was conducted. This solution allows optimising calculation, avoiding the models that provide worse results at the very beginning.

4.4. Statistical Analysis

In order to compare the variance of values (for 20 features) from three DSST parts, one-way ANOVA analysis was used. The Kolmogorov–Smirnov test (K-S test) and the Levene test were used to test the assumptions of the ANOVA analysis. The K-S test was performed to verify that variables had a normal distribution, the Levene test was used to check that the variance of the data from three parts of the DSST test was equal. Finally, the Tukey's honest significant difference test (Tukey's HSD) post-hoc test was performed to identify which pairs of the DSST test parts had statistically significant differences. The analysis was carried out in MATLAB 2020a software using Statistical and Machine Learning Toolbox.

5. Results

5.1. Classification Results

The three-class subject-independent classification was conducted and the F1 score of a selected classifier for a complete feature set is presented in Table 1. The following classifiers were applied: SVM with linear, quadratic, and cubic kernels, Logistic Regression, Decision Tree, kNN, Multilayer Perceptron, and Random Forest. The learning process including a train–test cycle was repeated 200 times in order to ensure the stability of results (the number was obtained empirically). Train and test datasets were formed randomly for every repetition. In order to ensure the methodological correctness of the experiments, the procedure of feature selection was performed on the training set independently in every repetition. Tables 2 and 3 present the results of numerical experiments: the main metrics allowing to assess the quality of classification: recall, precision, F1, accuracy, and ROC AUC.

Table 2. Main classification performance measures obtained for a complete feature set.

Classifier	Recall	Precision	F1 Score	Accuracy	ROC AUC
SVM linear	**0.94 ± 0.05**	**0.95 ± 0.05**	**0.94 ± 0.05**	**0.94 ± 0.05**	**0.99 ± 0.02**
SVM quadratic	0.71 ± 0.10	0.77 ± 0.10	0.71 ± 0.10	0.71 ± 0.10	0.85 ± 0.07
SVM cubic	0.90 ± 0.07	0.92 ± 0.06	0.90 ± 0.07	0.90 ± 0.07	0.98 ± 0.03
Log regression	**0.95 ± 0.05**	**0.96 ± 0.04**	**0.95 ± 0.05**	**0.95 ± 0.05**	**0.99 ± 0.02**
kNN	0.88 ± 0.07	0.90 ± 0.06	0.88 ± 0.07	0.88 ± 0.07	0.96 ± 0.04
Decision Tree	0.89 ± 0.07	0.92 ± 0.05	0.89 ± 0.07	0.89 ± 0.07	0.95 ± 0.04
Random Forest	**0.95 ± 0.05**	**0.96 ± 0.04**	**0.95 ± 0.05**	**0.95 ± 0.05**	**0.99 ± 0.02**
MLP	0.90 ± 0.07	0.92 ± 0.06	0.90 ± 0.07	0.90 ± 0.07	0.98 ± 0.03

The best classification performance results are in bold. The best mean values of separate performance metrics achieved by specific models are presented in the following way: mean ± standard deviation (feature number).

Table 3. Main classification performance measures obtained for a selected feature subset.

Classifier	Recall	Precision	F1 Score	Accuracy	ROC AUC
SVM linear	**0.97 ± 0.04 (5)**	**0.97 ± 0.03 (5)**	**0.97 ± 0.04 (5)**	**0.97 ± 0.04 (5)**	**0.99 ± 0.02 (5)**
SVM quadratic	0.92 ± 0.06 (8)	0.93 ± 0.05 (8)	0.92 ± 0.07 (8)	0.92 ± 0.06 (8)	0.98 ± 0.03 (8)
SVM cubic	0.94 ± 0.05 (8)	0.96 ± 0.04 (8)	0.94 ± 0.05 (8)	0.94 ± 0.05 (8)	0.99 ± 0.02 (6)
Log regression	**0.97 ± 0.04 (4)**	**0.97 ± 0.04 (4)**	**0.97 ± 0.04 (4)**	**0.97 ± 0.04 (4)**	**0.99 ± 0.01 (4)**
kNN	0.96 ± 0.05 (8)	0.96 ± 0.04 (8)	0.96 ± 0.05 (8)	0.96 ± 0.05 (8)	0.99 ± 0.02 (5)
Decision Tree	0.90 ± 0.07 (5)	0.92 ± 0.05 (5)	0.90 ± 0.07 (5)	0.90 ± 0.07 (5)	0.95 ± 0.05 (8)
Random Forest	0.95 ± 0.05 (7)	0.96 ± 0.03 (7)	0.95 ± 0.04 (7)	0.95 ± 0.05 (7)	0.99 ± 0.02 (7)
MLP	**0.97 ± 0.05 (5)**	**0.98 ± 0.04 (5)**	**0.97 ± 0.05 (5)**	**0.97 ± 0.05 (5)**	**0.99 ± 0.01 (5)**

Every cell contains a series of numbers, which has to be understood in the following way: mean ± standard deviation (number of features ensuring the best result). The best classification performance results are in bold. The best mean values of separate performance metrics achieved by specific models are presented in the following way: mean ± standard deviation (feature number).

All the measures, except for accuracy, were computed in a multiclass way; i.e., first, a measure was computed for the separate classes, and the final value was obtained as the mean of class-wise measure values with the weights proportional to the content of specific class samples in the test set.

Table 2 shows the values of classification performance measures obtained for the complete feature set; Table 3 presents the results obtained with the use of the feature selection procedure, based on elastic net, which is composed of logistic regression with linearly combined L1 and L2 regularization.

As it may be noticed, the best classification quality was achieved by MLP, SVM with linear kernel, and Logistic Regression. The mentioned models required correspondingly five, five, and four features per class. So, the total number of features was as high as seven, seven, and six—the numbers are obtained as the number of common features in the first five and four rows of Table 4, presenting the features ranked with respect to their importance for classifying a sample as an instance of the particular classes. Figure 4 presents the weights of the specific features for particular classes, which were marked with the color and shape of the markers. The whiskers represent the standard deviation of the values. Figure 4 presents the weights of the specific features, so, in addition to the rankings, the importance measures can be compared as well, so one can observe not only which features are more important but also quantitatively examine the difference of importance measures; e.g., it can be noticed that some features are significantly more important: mean amplitude of saccades, number of fixations, response number as well as mean response time.

Table 4. Separate class feature rankings obtained by interpreting the weights of the LogReg model with elastic net regularization.

No.	Low	Medium	High
1	mean amplitude of saccades	mean response time	response number
2	mean response time	response number	fixation number
3	standard deviation of fixation duration	mean amplitude of saccades	saccade number
4	fixation number	standard deviation of fixation duration	mean amplitude of saccades
5	standard deviation of right pupil diameter	fixation number	mean duration of fixation
6	saccade number	number of errors	maximum saccade amplitude
7	number of errors	mean duration of fixation	mean response time
8	mean of right pupil diameter	saccade number	standard deviation of right pupil diameter
9	mean duration of fixation	mean duration of saccades	mean duration of saccades
10	mean duration of saccades	standard deviation of right pupil diameter	maximum fixation duration
11	mean of blink duration	mean of right pupil diameter	number of errors
12	minimum fixation duration	maximum fixation duration	mean of right pupil diameter
13	response number	maximum saccade amplitude	blink number
14	standard deviation of saccades amplitude	minimum fixation duration	mean of blink duration
15	mean of left pupil diameter	standard deviation of saccades amplitude	standard deviation of left pupil diameter
16	standard deviation of left pupil diameter	mean of left pupil diameter	standard deviation of fixation duration
17	maximum fixation duration	blink number	standard deviation of saccades amplitude
18	maximum saccade amplitude	mean of blink duration	minimum fixation duration
19	blink number	standard deviation of left pupil diameter	mean of left pupil diameter
20	minimum saccade amplitude	minimum saccade amplitude	minimum saccade amplitude

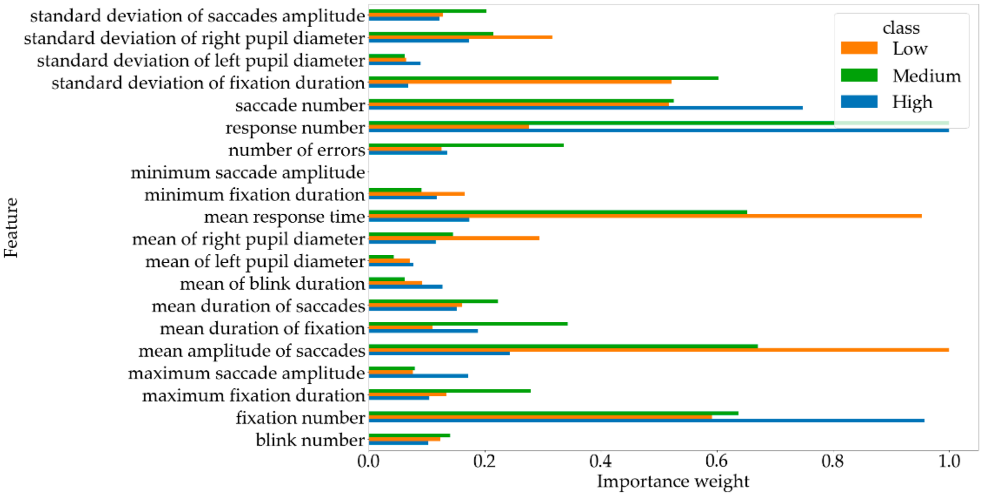


Figure 4. Separate feature importance weights.

Figures 5–7 show the dependence between the number of features applied and the results of the models, which achieved the highest values of F1 score after feature selection. The vertical lines indicate standard deviation of the results. As may be observed in the figures, the procedure of feature selection influences classification performance positively: all the plots presented in Figures 5–7 demonstrate a distinctive maximum for specific feature number and decrease of the F1 measure for a higher feature number (for SVM and logistic regression). Figures 5–7 present quantitative evidence of positive effect produced by feature selection and demonstrate that together with the decrease of computing complexity, an improvement in classification quality is achieved by selecting proper features.

Tables 5–7 present the mean confusion matrices of the models that provided the best performance, namely: SVM with a linear kernel, logistic regression, MLP. For example, for the SVM classifier, the observations from class 1 and class 2 are more similar to each other than the observations from class 3. This can be seen in the mean confusion matrix. All the observations of class 3 from the test dataset were classified correctly. On average, 5.62 observations of the first class were classified in the proper way and 0.38 observations from the first class were classified as second-class observations. The mean of 5.77 observation from the second class were classified in a proper way, and 0.23 observations from the second class were classified as first class. Only 0.005 observations on average were classified as the observations of the second class being the observations of the third class, while no observations of the third class were classified as first class.

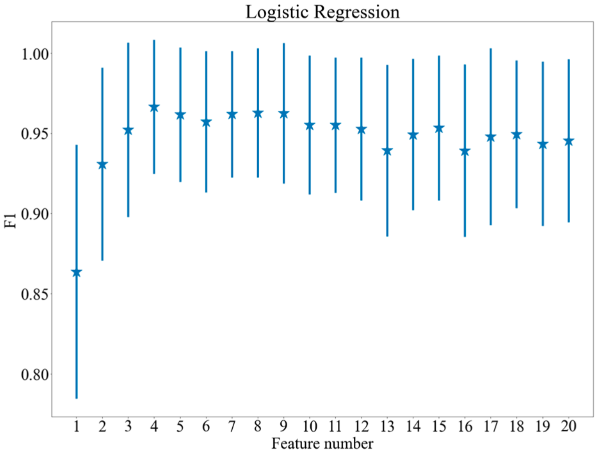


Figure 5. F1 scores for various feature numbers. Logistic Regression.

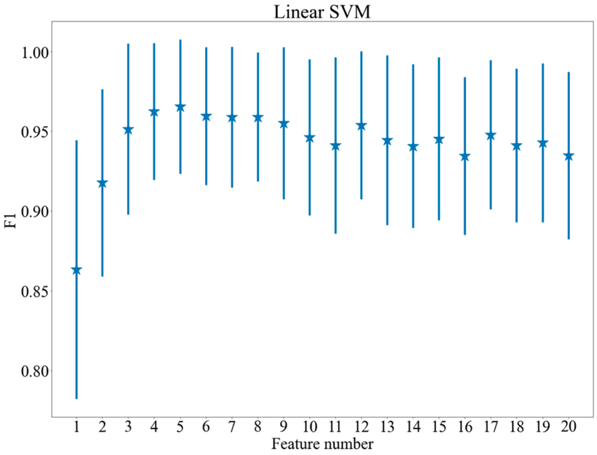


Figure 6. F1 scores for various feature numbers. Linear Support Vector Machines (Linear SVM).

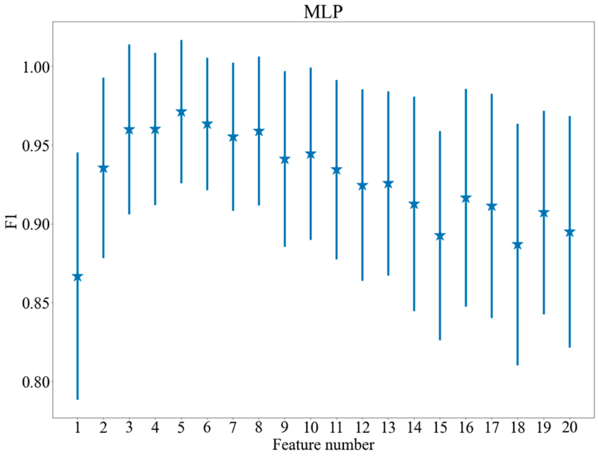


Figure 7. F1 scores for various feature numbers. Multilayer Perceptron (MLP).

Table 5. Mean confusion matrix for SVM classifier with linear kernel.

		True		
		Class 1	**Class 2**	**Class 3**
Predicted	**Class 1**	5.62	0.38	0
	Class 2	0.23	5.77	0
	Class 3	0	0.005	5.995

Table 6. Mean confusion matrix for Logistic Regression classifier with linear kernel.

		True		
		Class 1	**Class 2**	**Class 3**
Predicted	**Class 1**	5.64	0.36	0
	Class 2	0.22	5.78	0
	Class 3	0	0.005	5.995

Table 7. Mean confusion matrix for MLP.

		True		
		Class 1	**Class 2**	**Class 3**
Predicted	**Class 1**	5.64	0.36	0
	Class 2	0.16	5.84	0
	Class 3	0	0	6

5.2. Statistical Results

The K-S test revealed three features with non-normal distribution (minimum fixation duration, minimum saccade amplitude, and number of blinks). The Levene test found five features with unequal variances (number of fixations, minimum fixation duration, number of saccades, minimum saccade amplitude, number of blinks). As a result, the ANOVA analysis was performed for 15 features. It found seven features for which differences between their mean values were significant (*p*-value 0.05). The Tukey's HSD post-hoc test identified pairs of DSST parts which presented statistically significant differences.

Table 8 presents significant results (*p*-value < 0.05) of the ANOVA analysis and a post hoc test for selected features. Significant differences were observed mainly between class 1 and class 2, as well as between class 2 and class 3. Only one feature, maximum saccade amplitude, presented significant differences between class 2 and class 3 (*p*-value = 0.046).

Table 8. The results of one-way ANOVA analysis.

	ANOVA	**Post-hoc Test**		
Features	***p*-Value**	***p*-Value** **Class 1–Class 2**	***p*-Value** **Class 1–Class 3**	***p*-Value** **Class 2–Class 3**
mean response time	<0.001	<0.001	<0.001	0.069
standard deviation of fixation duration	0.002	0.003	0.008	0.947
maximum fixation duration	0.009	0.011	0.04	0.877
maximum saccade amplitude	0.002	0.411	0.001	0.046
mean saccade amplitude	<0.001	<0.001	<0.001	0.088
standard deviation of left pupil diameter	0.006	0.016	0.011	0.993
standard deviation of right pupil diameter	0.023	0.069	0.029	0.934

Figure 8 shows the comparison of mean values of the features that demonstrated the significant differences for various classes of cognitive workload.

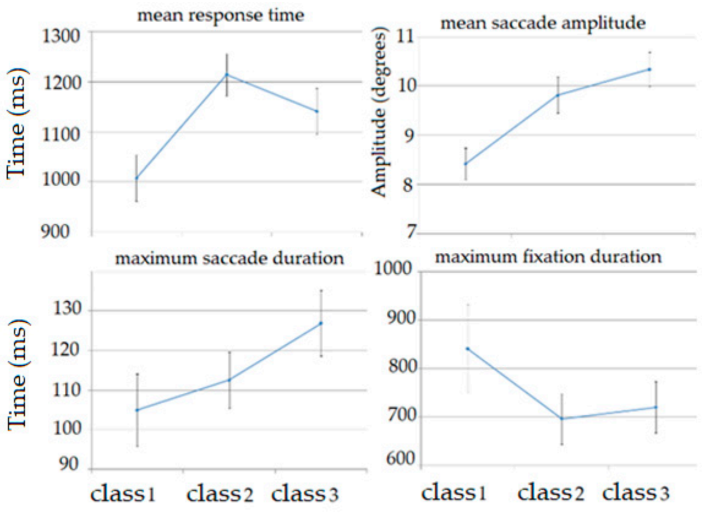


Figure 8. The comparison of mean values from the three-part digit symbol substitution test (DSST) for selected features.

5.3. Cognitive Factors Analysis

The topic of cognitive workload assessment also requires cognitive factor analysis. The term cognitive factors refers to characteristics of the person that affect his/her performance and learning effectiveness. Cognitive factors include such functions as memory, reasoning, and attention. Analysis of eye activity as well as DSST features allows estimating the cognitive factors. The procedure of feature selection run with the use of logistic regression model enabled obtaining the most valuable features, showing that the most important feature subsets are fixation and saccade-related features. Those kinds of features correspond to the intensity of eye movement, which shows a higher degree of attention during the performance of more complicated tasks, which is demonstrated especially well in Figure 8: in maximum fixation duration, we see that with the increase of task level from low to medium, the maximal duration of fixation decreases, which shows that the gaze of the examined person on average stays shorter in one position, which can be more evidence of higher attention in these tasks; the rest of the features presented in Figure 8 also support this thesis.

6. Discussion

The aim of the study was to perform the multiclass subject-independent classification of cognitive workload level with both interpretable and noninterpretable machine learning models. A three-class subject-independent classification was performed on the basis of the dataset containing eye-tracking and user performance data. The study assessed mental fatigue with features based on eye-tracking and DSST test results. The data were gathered in a case study of three versions of the computerised Digit Symbol Substitution Test (DSST). An interpretable machine learning model was used for selecting the most valuable features, which allowed improving the result of classification and obtaining insights that enabled understanding the process of mental fatigue.

Eight machine learning models were built and compared. It is a common approach to run an initial stage of classification tests with the use of multiple classifier models in order to find the most promising ones for further analysis. The aim of our work was not only to obtain the highest possible result of the classification but also to gain the feature interpretability especially in the case of subject-independent classification. The initial stage was provided by SVM with linear, quadratic, and cubic kernel, Logistic Regression, kNN, Decision Tree, MLP, and Random Forest. The learning process for each model was repeated 200 times.

Logistic Regression, chosen in the study as a feature selection tool, is commonly applied as an interpretable machine learning model, as it assigns specific weights to separate features, which allows assessing their importance quantitatively and hence creating a ranking. Logistic Regression was applied for feature selection, as it is commonly used in the cases requiring interpretable machine learning. It allowed improving the result from 0.95 to 0.97 (SVM with linear kernel) using only five features per binary classifier out of 20. Logistic Regression demonstrated a similar result with an even lower feature number (4). MLP also ensured high performance, which required five features. Based on the obtained results, it may be noticed that the most reasonable solution is to use an SVM/logistic regression model in the present problem.

The DSST test used in the study was a computerised version of the classical paper-and-pencil test. This cognitive test was chosen in the study as it was sensitive to changes in cognitive functioning, and its performance correlates with the ability to accomplish everyday tasks. Although originally, the test was designed to measure cognitive dysfunction, we decided to apply it in the study, as it is easy to use and it engages the memory and concentration of the participant. A preliminary pilot study was applied in order to define the number of symbols and duration of the test. The original version of the test assumes nine different symbols, which were generated in a random and repeated way to be assigned over a 90 s period. We decided to use this setting for the medium level of cognitive workload, whereas the settings for the low and hard level of cognitive workload were defined in an interview with the pilot study participants. However, the limitation of such an approach is that the level of cognitive workload is set permanently for all participants independently of their abilities and IQ level. It is worth noting that there are some methods dedicated for the evaluation of mental fatigue of particular participants. These methods take advantage of the capacity models (e.g., the Rasch and strain–stress model) and allow adjusting the cognitive workload assessment to a specific person, taking into consideration his/her mental abilities. On the other hand, these methods are based on surveys (e.g., NASA-TLX scale) and gather subjective assessment of the participant, which might also be blurred with different factors. Moreover, there are many various aspects affecting the subjective cognitive workload assessment, which are hard to consider and explain. A problem that can appear here is a more complicated structure of the experiment due to a possible lack of balance between the classes in this case. Nevertheless, supplementing the study with a mental fatigue evaluation of each participant might provide new insights into the analysis results and will be performed as a future work.

The paper presents the classification process based on eye activity gathered with the eye-tracking technique. This dataset is supported with features retrieved directly from the DSST application. The study includes general eye activity features related to visual fixations and saccadic movements, blinking, and pupil characteristics. Information about location indicating where participants were looking was not included. Such data might be used to map eye position onto the visual objects displayed on the screen; however, it is doubtful whether these results could increase the accuracy of the classification. On the other hand, such data could provide information about the test results, although these results are obtained directly from the DSST application.

Classification results and the feature ranking obtained in the study strongly suggest that the performed tasks have a systematic influence on eye movements. The results prove the relation between the participants' engagement in the task combined with their cognitive state and their eye activity. The results give insights into understanding the dependence between eye movements and cognitive factors. However, further research is necessary to explain these dependences among different participant groups and different stimulus types. Although the results of our study suggest that eye movement-related features might be applied in the process of cognitive state assessment, there is a possibility that the type of tasks, graphical interface, or even initial mental state of the participant might affect the results. However, both sets of features applied in the study (eye movement-related features and DSST test-related features) are objective measures independent of the subjective assessment of individual participants. What is more, the eye tracking-based features chosen in the study are a natural type of response gathered from a non-invasive source and obtained without any training or additional activity. A limitation of the work is the relatively small number of participants. Even if the group of 29 participants turned out to be sufficient for performing the statistical analysis, further studies are definitely necessary to confirm the results over a broader group of participants. What is more, we also plan to check the influence of the task order by randomising it in the experiment.

The highest F1 rate was achieved for SVM with a linear kernel after performing feature selection. Three binary classification models were used for distinguishing between three classes, each of them used an individual six-feature set obtained from the ranking built in the stage of feature selection (the rankings are presented in Table 4). The union of the binary model feature sets contains seven features: mean saccade amplitude, mean response time, fixation number, standard deviation of fixation duration, saccade number, response number, and mean duration of fixation. It might be noticed that the majority of the selected features are related to fixation (standard deviation of fixation duration, fixation number, mean duration of fixation) and saccadic eye movements (mean saccade amplitude, saccade number). It worth noting that only a standard deviation of fixation duration was important among the group containing an additional set of features (including standard deviation of saccades amplitude and fixation duration as well as maximum and minimum saccade amplitude and fixation duration). Such phenomenon could be related to the high variability of fixation duration among particular cognitive workload levels caused by possible physical eye fatigue, although this issue needs further investigation. No blink or pupillarity-related features were included in the set of seven selected features. This result suggests that the most commonly used eye-tracking features related to fixation and saccades are the most informative. The obtained results show that the cognitive workload level is related to the number of saccadic movements and fixation duration. Surprisingly, the analysis has shown low importance of blink-related features, which happened to be low discriminative in terms of cognitive workload. Additionally, according to previous expectations, the features related to DSST results also proved to be important for distinguishing between cognitive workload levels, which is quite intuitive; i.e., a more complicated task requires more time to solve and the probability of error is higher. The complete feature rankings are presented in Table 4. The feature rankings can be analytically understood and interpreted; for example, the analysis revealed that the response number was the most important feature for distinguishing the high level of cognitive workload, which is quite intuitive, as complicated tasks require a longer time to solve. More valuable information was related to the next features: fixation number and saccade number, which is also intuitive: participants tend to move their gaze more rapidly during solving more complex tasks. The results presented in Figure 5 allow comparing the scale of separate feature importance values, showing that some features are significantly more important than others, e.g., mean saccade amplitude, which shows that while solving easier tasks, the participants could move their gaze wider, producing longer saccades.

It has to be taken into account that differences in education, performed job, experience, and age can cause complication in the analysis, and results might differ between these groups. Thus, due to the approximate mental homogeneity of the examined group in the present research, the results show the relation between cognitive workload level and eye activity especially adapted to analytical minds of students of technical specialties.

The statistical analysis was based on the ANOVA procedure. Statistically significant differences for all classes were revealed for the maximum saccade amplitude feature, but the difference was observed only in the following pairs: class 1 and class 3, class 2 and class 3. The most significant differences were observed between class 1, which corresponds to a low cognitive workload level, and class 3, which is related to a high cognitive workload level, as they are the farthest from each other. However, the differences between average values of the mean response time feature was calculated for class 1 and class 2 but not for class 1 and class 3. It could be explained by the fact that the participants had been better acquainted with the application, so they answered questions faster, despite the fact that the third part of the experiment was more difficult. Presumably, there is no statistical significant difference between class 2 and class 3 for mean response time because the second and third part of the experiment included the same number of elements. What is more, a statistically significant difference has been found only for the maximum saccade amplitude feature between class 2 and class 3. This might be explained by the fact that the third part is the most advanced, and then, the participant saccades had the highest amplitude.

7. Conclusions

The specific aim of the paper was to conduct a deeper analysis of the features related to the classification of particular cognitive workload levels. It is an important task from the point of view of understanding the influence of cognitive workload level and mental fatigue on the brain cognitive process.

Interpretable machine learning allows to understand which features provide the most valuable information about the examined process. Generally speaking, it is more profitable to understand the reasons behind the decision taken by a machine learning model rather than using it as a black box. Understanding major processes beneath phenomena of interest allows us to build more robust models and perform more effective monitoring of cognitive workload.

For example, in the presented case, due to the interpretable machine learning model, we know which features to focus on and which ones can be neglected, thus allowing lowering the computing cost and obtaining better results. In practice, it is more reasonable to use as few features as possible, since this approach requires less processing power and is more robust.

As a general conclusion, the authors may state that the paper can serve as an example for researchers seeking ideas and techniques to investigate the relationships between mental fatigue and various biomedical measures.

Author Contributions: Conceptualization, M.P.-W. and M.K.; methodology, M.P.-W., M.K., M.T.; validation, M.P.-W., M.K., M.T.; formal analysis, M.P.-W., M.K., M.T.; resources, M.K.; data curation, M.K.; writing—original draft preparation, M.P.-W., M.K., M.T.; writing—review and editing, M.P.-W., M.K., M.T.; visualisation, M.K., M.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Ethics Committee of Lublin University of Technology (Approval ID 2/2018 from 19 October 2018).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: We would like to thank Martyna Wawrzyk for her help in data registration and analysis.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gevins, A.; Smith, M.E.; McEvoy, L.; Yu, D. High-resolution EEG mapping of cortical activation related to working memory: Effects of task difficulty, type of processing, and practice. *Cereb. Cortex* **1997**, *7*, 374–385. [CrossRef]
2. Qi, P.; Ru, H.; Sun, Y.; Zhang, X.; Zhou, T.; Tian, Y.; Thakor, N.; Bezerianos, A.; Li, J.; Sun, Y. Neural Mechanisms of Mental Fatigue Revisited: New Insights from the Brain Connectome. *Engineering* **2019**, *5*, 276–286. [CrossRef]
3. Gavelin, H.M.; Neely, A.S.; Dunås, T.; Eskilsson, T.; Järvholm, L.S.; Boraxbekk, C.-J. Mental fatigue in stress-related exhaustion disorder: Structural brain correlates, clinical characteristics and relations with cognitive functioning. *NeuroImage Clin.* **2020**, *27*, 102337. [CrossRef]
4. Grier, R.A.; Warm, J.S.; Dember, W.N.; Matthews, G.; Galinsky, T.L.; Szalma, J.L.; Parasuraman, R. The Vigilance Decrement Reflects Limitations in Effortful Attention, Not Mindlessness. *Hum. Factors J. Hum. Factors Ergon. Soc.* **2003**, *45*, 349–359. [CrossRef]
5. Van der Linden, D.; Eling, P. Mental fatigue disturbs local processing more than global processing. *Psychol. Research* **2006**, *70*, 395–402. [CrossRef] [PubMed]
6. Mackworth, N.H. The Breakdown of Vigilance during Prolonged Visual Search. *Q. J. Exp. Psychol.* **1948**, *1*, 6–21. [CrossRef]
7. Marquart, G.; Cabrall, C.; de Winter, J. Review of eye-related measures of drivers' mental workload. *Proc. Manuf.* **2015**, *3*, 2854–2861. [CrossRef]
8. Miller, S. *Workload Measures. National Advanced Driving Simulator*; University of Iowa Press: Iowa City, IA, USA, 2001.
9. Thummar, S.; Kalariya, V. A real time driver fatigue system based on eye gaze detection. *Int. J. Eng. Res. Gen. Sci.* **2015**, *3*, 105–110.
10. Wobrock, D.; Frey, J.; Graeff, D.; De La Rivière, J.-B.; Castet, J.; Lotte, F. Continuous Mental Effort Evaluation During 3D Object Manipulation Tasks Based on Brain and Physiological Signals. In Proceedings of the IFIP Conference on Human-Computer Interaction, Bamberg, Germany, 14–18 September 2015; Springer Nature: Cham, Switzerland, 2015; Volume 9296, pp. 472–487.
11. Son, J.; Oh, H.; Park, M. Identification of driver cognitive workload using support vector machines with driving performance, physiology and eye movement in a driving simulator. *Int. J. Precis. Eng. Manuf.* **2013**, *14*, 1321–1327. [CrossRef]
12. Matthews, G.; Reinerman-Jones, L.E.; Barber, D.J.; Abich IV, J. The psychometrics of mental workload: Multiple measures are sensitive but divergent. *Hum. Factors* **2015**, *57*, 125–143. [CrossRef] [PubMed]
13. Henderson, J.M.; Shinkareva, S.V.; Wang, J.; Luke, S.G.; Olejarczyk, J. Predicting Cognitive State from Eye Movements. *PLoS ONE* **2013**, *8*, e64937. [CrossRef] [PubMed]
14. Benfatto, M.N.; Öqvist Seimyr, G.; Ygge, J.; Pansell, T.; Rydberg, A.; Jacobson, C. Screening for Dyslexia Using Eye Tracking during Reading. *PLoS ONE* **2016**, *11*, e0165508. [CrossRef]

15. Mark, J.; Curtin, A.; Kraft, A.; Sands, T.; Casebeer, W.D.; Ziegler, M.; Ayaz, H. Eye Tracking-Based Workload and Performance Assessment for Skill Acquisition. In *Advances in Intelligent Systems and Computing*; Springer Nature: Cham, Switzerland, 2019; Volume 953, pp. 129–141.
16. Coco, M.I.; Keller, F. Classification of visual and linguistic tasks using eye-movement features. *J. Vis.* **2014**, *14*, 11. [CrossRef]
17. Lobo, J.L.; Del Ser, J.; De Simone, F.; Presta, R.; Collina, S.; Moravek, Z. Cognitive workload classification using eye-tracking and EEG data. In Proceedings of the International Conference on Human-Computer Interaction in Aerospace, ACM 2016, Paris, France, 14–16 September 2016; pp. 1–8.
18. Chen, J.; Wang, H.; Wang, Q.; Hua, C. Exploring the fatigue affecting electroencephalography based functional brain networks during real driving in young males. *J. Neuropsychol.* **2019**, *129*, 200–211. [CrossRef] [PubMed]
19. Nuamah, J.K.; Seong, Y. Support vector machine (SVM) classification of cognitive tasks based on electroencephalography (EEG) engagement index. *Br. Comput. Interf.* **2017**, *5*, 1–12. [CrossRef]
20. Chen, L.-L.; Zhao, Y.; Ye, P.-F.; Zhang, J.; Zou, J.-Z. Detecting driving stress in physiological signals based on multimodal feature analysis and kernel classifiers. *Expert Syst. Appl.* **2017**, *85*, 279–291. [CrossRef]
21. Khushaba, R.N.; Kodagoda, S.; Lal, S.; Dissanayake, G. Driver Drowsiness Classification Using Fuzzy Wavelet-Packet-Based Feature-Extraction Algorithm. *IEEE Trans. Biomed. Eng.* **2011**, *58*, 121–131. [CrossRef]
22. Atasoy, H.; Yildirim, E. Classification of Verbal and Quantitative Mental Tasks Using Phase Locking Values between EEG Signals. *Int. J. Signal Process. Image Process. Pattern Recognit.* **2016**, *9*, 383–390. [CrossRef]
23. Zarjam, P.; Epps, J.; Lovell, N.H. Beyond Subjective Self-Rating: EEG Signal Classification of Cognitive Workload. *IEEE Trans. Auton. Ment. Dev.* **2015**, *7*, 301–310. [CrossRef]
24. Magnusdottir, E.H.; Johannsdottir, K.R.; Bean, C.; Olafsson, B.; Gudnason, J. Cognitive workload classification using cardiovascular measures and dynamic features. In Proceedings of the 8th IEEE International Conference on Cognitive Infocommunications (CogInfo-Com), Debrecen, Hungary, 11–14 September 2017; pp. 351–356.
25. Spüler, M.; Walter, C.; Rosenstiel, W.; Gerjets, P.; Moeller, K.; Klein, E. EEG-based prediction of cognitive workload induced by arithmetic: A step towards online adaptation in numerical learning. *ZDM* **2016**, *48*, 267–278. [CrossRef]
26. Laine, T.; Bauer, K.; Lanning, J.; Russell, C.; Wilson, G. Selection of input features across subjects for classifying crewmember workload using artificial neural networks. *IEEE Trans. Syst. Man Cybern. Part A Syst. Hum.* **2002**, *32*, 691–704. [CrossRef]
27. Wang, Z.; Hope, R.M.; Wang, Z.; Ji, Z.; Gray, W.D. Cross-subject workload classification with a hierarchical Bayes model. *NeuroImage* **2012**, *59*, 64–69. [CrossRef]
28. Walter, C.; Wolter, P.; Rosenstiel, W.; Bogdan, M.; Spüler, M. Towards cross-subject workload prediction. In Proceedings of the 6th International Brain-Computer Interface Conference, Graz, Austria, 16–19 September 2014.
29. Fazli, S.; Mehnert, J.; Steinbrink, J.; Curio, G.; Villringer, A.; Müller, K.-R.; Blankertz, B. Enhanced performance by a hybrid NIRS–EEG brain computer interface. *NeuroImage* **2012**, *59*, 519–529. [CrossRef]
30. Thodoroff, P.; Pineau, J.; Lim, A. Learning robust features using deep learning for automatic seizure detection. In Proceedings of the Machine Learning for Healthcare Conference, Los Angeles, CA, USA, 19–20 August 2016; pp. 178–190.
31. Boake, C. From the Binet–Simon to the Wechsler–Bellevue: Tracing the History of Intelligence Testing. *J. Clin. Exp. Neuropsychol.* **2002**, *24*, 383–405. [CrossRef]
32. Sicard, V.; Moore, R.D.; Ellemberg, D. Sensitivity of the Cogstate Test Battery for Detecting Prolonged Cognitive Alterations Stemming From Sport-Related Concussions. *Clin. J. Sport Med.* **2019**, *29*, 62–68. [CrossRef]
33. Cook, N.; Kim, J.U.; Pasha, Y.; Crossey, M.M.; Schembri, A.J.; Harel, B.T.; Kimhofer, T.; Taylor-Robinson, S.D. A pilot evaluation of a computer-based psychometric test battery designed to detect impairment in patients with cirrhosis. *Int. J. Gen. Med.* **2017**, *10*, 281–289. [CrossRef] [PubMed]
34. Jaeger, J. Digit symbol substitution test: The case for sensitivity over specificity in neuropsychological testing. *J. Clin. Psychopharm.* **2018**, *38*, 513. [CrossRef]
35. McKendrick, R.; Feest, B.; Harwood, A.; Falcone, B. Theories and Methods for Labeling Cognitive Workload: Classification and Transfer Learning. *Front. Hum. Neurosci.* **2019**, *13*, 295. [CrossRef] [PubMed]
36. Işbilir, E.; Çakır, M.P.; Acartürk, C.; Tekerek, A. Şimşek Towards a Multimodal Model of Cognitive Workload Through Synchronous Optical Brain Imaging and Eye Tracking Measures. *Front. Hum. Neurosci.* **2019**, *13*, 375. [CrossRef] [PubMed]
37. Ziegler, M.D.; Kraft, A.; Krein, M.; Lo, L.-C.; Hatfield, B.; Casebeer, W.; Russell, B. Sensing and Assessing Cognitive Workload Across Multiple Tasks. In Proceedings of the International Conference on Augmented Cognition, Toronto, ON, Canada, 17–22 July 2016; Springer Nature: Chan, Switzerland, 2016; pp. 440–450.
38. Almogbel, M.A.; Dang, A.H.; Kameyama, W. EEG-signals based cognitive workload detection of vehicle driver using deep learning. In Proceedings of the 2018 20th International Conference on Advanced Communication Technology (ICACT), Chuncheon, Korea, 11–14 February 2018; pp. 256–259.
39. Almogbel, M.A.; Dang, A.H.; Kameyama, W. Cognitive Workload Detection from Raw EEG-Signals of Vehicle Driver using Deep Learning. In Proceedings of the 2019 21st International Conference on Advanced Communication Technology (ICACT), PyeongChang, Korea, 17–20 February 2019; pp. 1–6.
40. Hefron, R.; Borghetti, B.J.; Kabban, C.M.S.; Christensen, J.C.; Estepp, J. Cross-Participant EEG-Based Assessment of Cognitive Workload Using Multi-Path Convolutional Recurrent Neural Networks. *Sensors* **2018**, *18*, 1339. [CrossRef]

41. Zarjam, P.; Epps, J.; Chen, F.; Lovell, N.H. Estimating cognitive workload using wavelet entropy-based features during an arithmetic task. *Comput. Biol. Med.* **2013**, *43*, 2186–2195. [CrossRef]
42. Appel, T.; Scharinger, C.; Gerjets, P.; Kasneci, E. Cross-subject workload classification using pupil-related measures. In Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications, Warsaw, Poland, 14–17 June 2018; pp. 1–8.
43. Hajinoroozi, M.; Mao, Z.; Jung, T.P.; Lin, C.T.; Huang, Y. EEG-based prediction of driver's cognitive performance by deep convolutional neural network. *Signal Proc. Imag. Commun.* **2016**, *47*, 549–555. [CrossRef]
44. Bozkir, E.; Geisler, D.; Kasneci, E. Person Independent, Privacy Preserving, and Real Time Assessment of Cognitive Load using Eye Tracking in a Virtual Reality Setup. In Proceedings of the 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), Osaka, Japan, 23–27 March 2019; pp. 1834–1837.
45. Fridman, L.; Reimer, B.; Mehler, B.; Freeman, W.T. Cognitive Load Estimation in the Wild. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Montreal, QC, Canada, 21–26 April 2018; p. 652.
46. Yamada, Y.; Kobayashi, M. Detecting mental fatigue from eye-tracking data gathered while watching video: Evaluation in younger and older adults. *Artif. Intell. Med.* **2018**, *91*, 39–48. [CrossRef]
47. Jimenez-Guarneros, M.; Gómez-Gil, P. Cross-subject classification of cognitive loads using a recurrent-residual deep network. In Proceedings of the 2017 IEEE Symposium Series on Computational Intelligence (SSCI), Honolulu, HI, USA, 27 November–1 December 2017; pp. 1–7. [CrossRef]
48. Appel, T.; Sevcenko, N.; Wortha, F.; Tsarava, K.; Moeller, K.; Ninaus, M.; Kasneci, E.; Gerjets, P. Predicting Cognitive Load in an Emergency Simulation Based on Behavioral and Physiological Measures. In Proceedings of the 2019 International Conference on Multimodal Interaction, Suzhou, Jiangsu, China, 14–18 October 2019; pp. 154–163.
49. Jimnez-Guarneros, M.; Gomez-Gil, P. Custom Domain Adaptation: A new method for cross-subject, EEG-based cognitive load recognition. *IEEE Sign. Proc. Let.* **2020**, *27*, 750–754. [CrossRef]
50. Chen, S.; Epps, J.; Ruiz, N.; Chen, F. Eye activity as a measure of human mental effort in HCI. In Proceedings of the 16th international conference on Intelligent user interfaces, Palo Alto, CA, USA, 13–16 February 2011; pp. 315–318.
51. Tokuda, S.; Obinata, G.; Palmer, E.; Chaparro, A. Estimation of mental workload using saccadic eye movements in a free-viewing task. In Proceedings of the 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE Engineering in Medicine and Biology Society, Boston, MA, USA, 30 August–3 September 2011; pp. 4523–4529.
52. Tobii AB. Tobii Studio User's Manual. Available online: https://www.tobiipro.com/siteassets/tobii-pro/user-manuals/tobii-pro-studio-user-manual.pdf (accessed on 7 October 2020).
53. Rayner, K. Eye movements in reading and information processing: 20 years of research. *Psychol. Bull.* **1998**, *124*, 372–422. [CrossRef] [PubMed]
54. Hessels, R.S.; Niehorster, D.C.; Nyström, M.; Andersson, R.; Hooge, I.T.C. Is the eye-movement field confused about fixations and saccades? A survey among 124 researchers. *R. Soc. Open Sci.* **2018**, *5*, 180502. [CrossRef] [PubMed]
55. Salvucci, D.D.; Goldberg, J.H. Identifying fixations and saccades in eye-tracking protocols. In Proceedings of the 2000 Symposium on Eye Tracking Research & Applications, Palm Beach Gardens, FL, USA, 6–8 November 2000; pp. 71–78.
56. Barbato, G.; Ficca, G.; Muscettola, G.; Fichele, M.; Beatrice, M.; Rinaldi, F. Diurnal variation in spontaneous eye-blink rate. *Psychiatry Res.* **2000**, *93*, 145–151. [CrossRef]
57. Shishido, E.; Ogawa, S.; Miyata, S.; Yamamoto, M.; Inada, T.; Ozaki, N. Application of eye trackers for understanding mental disorders: Cases for schizophrenia and autism spectrum disorder. *Neuropsychopharmacol. Rep.* **2019**, *39*, 72–77. [CrossRef]
58. Olsen, A.; Matos, R. Identifying parameter values for an I-VT fixation filter suitable for handling data sampled with various sampling frequencies. In Proceedings of the Symposium on Eye Tracking Research and Applications, Santa Barbara, CA, USA, 28–30 March 2012; p. 317.
59. Kardan, O.; Berman, M.G.; Yourganov, G.; Schmidt, J.; Henderson, J.M. Classifying mental states from eye movements during scene viewing. *J. Exp. Psychol. Hum. Percept. Perform.* **2015**, *41*, 1502–1514. [CrossRef]
60. Dowiasch, S.; Marx, S.; Einhäuser, W.; Bremmer, F. Effects of aging on eye movements in the real world. *Front. Hum. Neurosci* **2015**, *9*, 1–12. [CrossRef] [PubMed]
61. Mould, M.S.; Foster, D.H.; Amano, K.; Oakley, J.P. A simple nonparametric method for classifying eye fixations. *Vis. Res.* **2012**, *57*, 18–25. [CrossRef]
62. Ryu, K.; Myung, R. Evaluation of mental workload with a combined measure based on physiological indices during a dual task of tracking and mental arithmetic. *Int. J. Ind. Ergon.* **2005**, *35*, 991–1009. [CrossRef]
63. Rozado, D.; Duenser, A.; Howell, B. Improving the Performance of an EEG-Based Motor Imagery Brain Computer Interface Using Task Evoked Changes in Pupil Diameter. *PLoS ONE* **2015**, *10*, e0121262. [CrossRef] [PubMed]
64. Partala, T.; Jokiniemi, M.; Surakka, V. Pupillary responses to emotionally provocative stimuli. In Proceedings of the 2000 Symposium on Eye Tracking Research & Applications, Palm Beach Gardens, FL, USA, 6–8 November 2000; pp. 123–129.
65. Tuszyńska-Bogucka, W.; Kwiatkowski, B.; Chmielewska, M.; Dzieńkowski, M.; Kocki, W.; Pełka, J.; Przesmycka, N.; Bogucki, J.; Galkowski, D. The effects of interior design on wellness—Eye tracking analysis in determining emotional experience of architectural space. A survey on a group of volunteers from the Lublin Region, Eastern Poland. *Ann. Agric. Environ. Med.* **2020**, *27*, 113–122. [CrossRef]

66. Mathôt, S.; Fabius, J.; Van Heusden, E.; Van Der Stigchel, S. Safe and sensible preprocessing and baseline correction of pupil-size data. *Behav. Res. Methods* **2018**, *50*, 94–106. [CrossRef] [PubMed]
67. Du, M.; Liu, N.; Hu, X. Techniques for interpretable machine learning. *Commun. ACM* **2019**, *63*, 68–77. [CrossRef]

6.3 On the Improvement of Eye Tracking-Based Cognitive Workload Estimation Using Aggregation Functions

1. Details

This article was written by Monika Kaczorowska, Paweł Karczmarek, Małgorzata Plechawska-Wójcik and Mikhail Tokovarov and published in Sensors in 2021, vol. 21, nr 13. The article is worth100 points according to the list of the Polish Ministry of Science and Higher Education.

2. Abstract

Cognitive workload, being a quantitative measure of mental effort, draws significant interest of researchers, as it allows to monitor the state of mental fatigue. Estimation of cognitive workload becomes especially important for job positions requiring outstanding engagement and responsibility, e.g., air-traffic dispatchers, pilots, car or train drivers. Cognitive workload estimation finds its applications also in the field of education material preparation. It allows to monitor the degree of difficulty for specific tasks enabling to adjust the level of education materials to typical abilities of students. In this study, we present the results of research conducted with the goal of examining the influence of various fuzzy or non-fuzzy aggregation functions upon the quality of cognitive workload estimation. Various classical machine learning models were successfully applied to the problem. The results of extensive in-depth experiments with over 2000 aggregation operators show the applicability of the approach based on the aggregation functions. Moreover, the approach based on the aggregation process allows for further improvement of classification results. A wide range of aggregation functions is considered and the results suggest that the combination of classical machine learning models and aggregation methods allows to achieve high quality of cognitive workload level recognition preserving low computational costs.

sensors

MDPI

Article

On the Improvement of Eye Tracking-Based Cognitive Workload Estimation Using Aggregation Functions

Monika Kaczorowska , Paweł Karczmarek, Małgorzata Plechawska-Wójcik * and Mikhail Tokovarov

Department of Computer Science, Lublin University of Technology, 20-618 Lublin, Poland;
m.kaczorowska@pollub.pl (M.K.); p.karczmarek@pollub.pl (P.K.); m.tokovarov@pollub.pl (M.T.)
* Correspondence: m.plechawska@pollub.pl

Abstract: Cognitive workload, being a quantitative measure of mental effort, draws significant interest of researchers, as it allows to monitor the state of mental fatigue. Estimation of cognitive workload becomes especially important for job positions requiring outstanding engagement and responsibility, e.g., air-traffic dispatchers, pilots, car or train drivers. Cognitive workload estimation finds its applications also in the field of education material preparation. It allows to monitor the difficulty degree for specific tasks enabling to adjust the level of education materials to typical abilities of students. In this study, we present the results of research conducted with the goal of examining the influence of various fuzzy or non-fuzzy aggregation functions upon the quality of cognitive workload estimation. Various classic machine learning models were successfully applied to the problem. The results of extensive in-depth experiments with over 2000 aggregation operators shows the applicability of the approach based on the aggregation functions. Moreover, the approach based on aggregation process allows for further improvement of classification results. A wide range of aggregation functions is considered and the results suggest that the combination of classical machine learning models and aggregation methods allows to achieve high quality of cognitive workload level recognition preserving low computational cost.

Keywords: aggregation; generalized Choquet integral; fuzzy measure; classical machine learning; cognitive workload

check for updates

Citation: Kaczorowska, M.; Karczmarek, P.; Plechawska-Wójcik, M.; Tokovarov, M. On the Improvement of Eye Tracking-Based Cognitive Workload Estimation Using Aggregation Functions. *Sensors* **2021**, *21*, 4542. https://doi.org/10.3390/s21134542

Academic Editors: Marco Porta, Pawel Kasprowski, Luca Lombardi and Piercarlo Dondi

Received: 21 May 2021
Accepted: 30 June 2021
Published: 2 July 2021

1. Introduction

Cognitive workload is understood as a mental effort necessary to perform a task [1]. It is a non-trivial process useful in explaining mental fatigue and its influence on the brain's cognitive system performance. Automatic categorizing and classification of cognitive workload levels is a subject of numerous research studies published recently. The classification of cognitive workload can be conducted in two ways: subject-dependent approach [2–4] and subject-independent approach [5,6]. Subject-independent approach, being more general, attracts greater attention of the researchers nowadays [7]. The literature review [8] also shows the examples of combined subject-dependent and subject-independent approaches. The most frequent case that can be found in the literature is binary classification problem: distinguishing between low and high levels of cognitive workload [9,10]. Besides the binary approach, papers dealing with three-way classification can be found. In that case, low, medium, and high levels of cognitive workload are considered [6,7,11]. Experiments involving multiclass classification are less common in the cognitive workload research [12,13]. The literature shows the reports of the results obtained with various classifiers, but the most popular among them are Support Vector Machine (SVM) [6,14,15], Linear Discriminant Analysis (LDA) [16], k-Nearest Neighbors (kNN) [11], and Random Forest [6]. In addition to classical recognition models, deep neural network-based approaches such as convolutional deep neural networks [9,17,18] are applied in the cognitive classification process. The reported results of accuracy are in the range of 50–80%. Classification of cognitive workload can be conducted on the basis of electroencephalographic (EEG) data [11], galvanic skin response (GSR) [19], or eye-tracking [20]. In [21,22], the authors use the fuzzy methods to effectively monitor the state of cognitive workload of an Unmanned Aerial Vehicle (UAV) operator. In [23], the authors successfully apply fuzzy cognitive mapping to analyze the pilots' decision during the flight.

It is worth recalling a few recent results. Fatimah and colleagues [24] published an article on the automatic detection of mental difficulty in arithmetic tasks on the basis of an EEG signal. The authors used a publicly available dataset from MIT PhysioNet, which contains recordings from 36 people. The arithmetic tasks performed by the respondents consisted of subtracting numbers. Based on the number of correct calculations per minute, the performed tasks were divided into two groups: easy and difficult. If the number of incorrect answers was not more than 20%, the tasks were considered easy, otherwise they were considered difficult. For 12 people, the tasks turned out to be easy, and for 24, the tasks were difficult. A two-class classification independent of the examined person was carried out: the main goal was to distinguish between low and high levels of cognitive load. The following classifiers were used: SVM, Decision Tree, and Quadratic Discriminant. Accuracy of the classification was calculated for each electrode separately and for each electrode divided into bands. The best results were achieved for the Quadratic Discriminant classifier, both with and without division into bands for a given electrode [24]. The best accuracy achieved with selected electrode and specific frequency band was as high as 97.2%. In [25], the authors conducted research aimed at detecting various mental states of the pilot such as distraction, workload, fatigue, and normal state. Various biosignals were used in the study: EEG, EKG, EDA, and EEA. Based on the signals collected from eight pilots, a four-class classification was carried out relating to distraction, workload, fatigue, and normal state. The authors presented the results of classification independent of the tested person for various classifiers, among others, for KNN, SVM, sLDA, LSTM, and their own proposed network for the EEG data separately, for the rest of the signals and for the combination of the EEG with the rest of the signals. The best results for the majority of classifiers were obtained for the data considering all signals. For the method proposed by the authors, based on the LSTM, the mean classification score was 85.2% (accuracy). In [26], the authors presented a model based on GALoRIS, thanks to which it is possible to identify high and low cognitive loads. The algorithm selects the features that correspond to low and high loads. The model was tested by the authors on the basis of the cognitive load data associated with driving. EEG data for the experiment were collected while driving the vehicle in the simulator. In addition, the authors used the NASA scale TLX and Instantaneous Self-Assessment (ISA), which enabled the subjective assessment of the individual and the vehicle performance measures (error level). The authors conducted a classification independent of the examined person and tested several classifiers in their research, the best result was achieved for the SVM classifier and was over 96%. Agnola and colleagues [27] dealt with a very interesting topic—the cognitive load in the context of using drones in search-and-rescue (SAR) missions. The authors used a simulator with which three levels of SAR-related cognitive bias were evoked. They used biological signals such as: ECG, skin temperature, respiration. The authors proposed a method of eliminating the extracted features using the following algorithms: eXtreme Gradient Boosting (XGBoost) and Shapley Additive exPlanations (SHAP). Experiment was carried out on 24 people who were asked to perform four activities: baseline, mapping activity, flying activity, flying and mapping activity simultaneously. As in the case of article [26], the authors used the NASA-TLX scale. The article presents the results of classification independent of the tested person, both two-class and three-class using such classifiers as kNN, Logistic regression, LDA, XGBoost, Random Forest. Two-class classification was used for distinguishing between low and high cognitive load. The authors obtained 80.2% accuracy for the two-class classification and 62.9% for the three-class classification using the XGBoost classifier with 24 features. In the paper [28], the authors presented a model that classifies the cognitive load based on the Long Short-Term Memory (LSTM) network and the Filter Bank Common Spatial

Pattern (FBCSP) based on EEG data. The authors conducted the two-class classification: arithmetical tasks and rest state; they achieved an accuracy of 87% with this model. In their research, the authors used a publicly available dataset, which contains data from 30 people performing arithmetic tasks.

The poor or unsatisfactory quality of some classifiers in various fields of application can be compensated by the use of appropriate operators aggregating the classification results returned by these classifiers separately or on the basis of an information fusion at the stage of the data preprocessing. The former way of finding the final ranking of classification results is intuitively appealing and typical for many fields of application such as sport competitions, risk analysis, decision-making, etc. These aggregation functions or operators are described in detail in many monographs [29–34] and papers [35–37]. In particular, typical classes of aggregation operators are means, triangular norms [38,39], ordinary weighted averaging operators [35,40], Choquet integral, and its generalizations [41–47] called pre-aggregation functions, etc. Comprehensive experimental studies, in particular, on an applications of aggregation operators and generalizations of Choquet integral to the face recognition problems were presented in [44,46,48], respectively.

The main goal of this study is to improve the results of eye activity and user performance-based cognitive workload level classification with the use of aggregation methods. For this purpose, we test and compare over 1000 classic aggregation operators and over 1000 pre-aggregation operators (so called generalized Choquet integrals) to determine the best one. The set of aggregation operators utilized in a series of thorough numerical experiments is built on the basis of above-mentioned monographs [29–34] and selected papers. We list the best aggregation functions and discuss the accuracies obtained for the typical classifiers such as Decision Tree, k-Nearest Neighbors, etc. The dataset used in the classification study contains eye-tracking and user performance data taken from 29 participants in the study of solving the computerized version of Digit Symbol Substitution Test (DSST).

The rest of the paper is structured as follows. Section 2 presents the description of the experiment procedure with detailed explanation of eyetracking-related aspects and data processing methods applied. Section 3 presents the utilized aggregation functions. Section 4 contains the presentation of the results obtained with individual classifiers as well as the recognition rates achieved with application of the presented aggregation functions. Section 5 concludes the paper and presents the future work directions.

2. Eyetracking

2.1. Research Procedure

The dataset containing eye activity and user performance data was gathered using the computerized version of the DSST test [49] developed for the purpose of this study. The idea of DSST test is to match displayed symbols to particular digits according to a key presented continuously on the screen (Figure 1). In the study, participants were asked to assign subsequent symbols to digits within the specified time. Symbols were generated randomly and with repetition. Participants were instructed to perform as many correct matches as possible within defined time. The time of single trial and the number of different symbols to be displayed were defined in the application settings. For the purpose of the study, three DSST parts were prepared; each of them corresponded to one cognitive workload level in the further analysis. Part 1 corresponding to the low level of cognitive workload, contained four different symbols, and the time was set to 90 s. Part 2 related to the medium level of cognitive workload, covered nine different symbols, and the time was also set to 90 s. Part 3 defined for the hard level of cognitive workload, covered nine different symbols, and the time was extended to 180 s. In all parts, participants were asked to perform as many matchings of subsequent symbols to digits as possible (in defined time). They were also instructed to perform matches as fast as possible. The settings were defined empirically based on the preliminary pilotage study. Each participant of the case study was asked to perform all three DSST parts. The experiment was preceded by short trial to familiarize participants with the application.

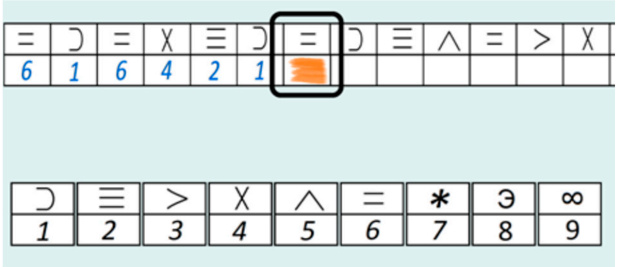


Figure 1. The interface of the application.

The experiment was performed in a laboratory room illuminated with standard fluorescent light. The eye activity data were gathered using Tobii Pro TX300 screen-based eye tracker (Tobii AB, Stockholm, Sweden), which was built into a monitor (23″ TFT monitor, 60 Hz) connected to the computer. Data were registered with the frequency of 300 Hz. Tobii Studio 3.2 software was used to design the experiment and export data. Each session was preceded by the 9-point calibration procedure.

Eye activities gathered in the experiment were related to such measures as fixations, saccades, blinks, and pupil size. Fixations are understood as the period of uptaking visual information, during which a participant holds eyes stable in a particular position. Saccades are understood as the rapid eye movement occurring between fixations. The dataset covered 20 selected features related to fixations (total number of fixations, mean duration of fixation, standard deviation of fixation duration, maximum fixation duration, minimum fixation duration), saccades (total number of saccades, mean duration of saccades, mean amplitude of saccades, standard deviation of saccade amplitude, maximum saccade amplitude, minimum saccade amplitude), blinks (total number of blinks, mean of blink duration), and pupillary response (mean of left pupil diameter, mean of right pupil diameter, standard deviation of left pupil diameter, standard deviation of right pupil diameter). Moreover, data related to DSST test results, i.e., number of errors, mean response time, and response number, were also included.

The experiment was conducted on a homogeneous group of 30 participants: 24 males, six females aged 20 to 24 (mean = 20.61 years, std. dev. = 1.54) recruited among healthy students of the BSc degree in computer science. The participants reported to have normal/corrected to normal vision and they were not under strong medicines. As the acceptable level of registered data activity was set to 90%, data from one participant were discarded from the further analysis due to their poor quality.

2.2. Data Processing

The data processing procedure was composed of six steps: data acquisition, data synchronization, feature extraction, feature normalization, feature selection, training, and testing classification models. The raw data were generated in the form of six files per single participant (two files (eyetracking data and DSST results) for each of three DSST parts). Owing to that fact, a synchronization procedure was needed. Finally, 87 observations were included in the output dataset (three observations representing three cognitive workload levels per single participant). In the feature extraction procedure, twenty independent features were obtained. Feature normalization was also performed to guarantee a uniform feature scale.

The ANOVA analysis was performed for 17 features. The K-S test and Levene test were previously performed to check assumptions of normality of distribution and equality of variance. In this process, three of 20 features (mean duration of saccades, minimum saccade amplitude, and mean of blink duration) were discarded from further analysis. The ANOVA analysis revealed 10 significant features (p-value 0.05), which were applied in classification process. The Tukey's HSD post-hoc test was applied in order to identify the pairs of DSST parts which differed significantly. Table 1 presents significant results (p-value < 0.05) of the ANOVA analysis.

Table 1. The results of one-way ANOVA analysis.

	ANOVA	Post-Hoc Test		
Features	***p*-Value**	***p*-Value Class 1–Class 2**	***p*-Value Class 1–Class 3**	***p*-Value Class 2–Class 3**
response number	<0.001	<0.001	<0.001	<0.001
mean response time	<0.001	<0.001	<0.001	0.69
total number of fixations	<0.001	0.36	<0.001	<0.001
standard deviation of fixation duration	0.002	0.003	0.008	0.95
maximum fixation duration	0.009	0.011	0.04	0.87
total number of saccades	<0.001	0.56	<0.001	<0.001
maximum saccade amplitude	0.002	0.41	0.001	0.046
mean saccade amplitude	<0.001	<0.001	0.09	<0.001
total number of blinks	0.015	0.99	0.003	0.003
standard deviation of pupil diameter (left)	0.005	0.016	0.012	0.99

The classification procedure was focused on assigning observations into one of the three classes: low, medium, and high level of cognitive workload. Various classification methods such as SVM, kNN, Decision Tree, Random Forest, Multilayer Perceptron (MLP), and Logistic Regression were applied. As the classification was performed using a subject-independent approach, the division into train and test datasets was done in such a way that a single participant could be used only in one dataset. The test dataset covered data from six participants, which corresponded to approximately 20% of the input dataset.

In order to investigate the influence of particular features of classification process, feature importance ranking was generated. Table 2 presents the features ranked with respect to their importance for classifying procedure. The results were obtained based on Logistic Regression model.

Table 2. Separate class feature rankings together with weights obtained by interpreting the weights of the Logistic Regression model.

No.	Low	Medium	High
1	mean saccade amplitude (1.0)	mean response time (1.0)	response number (1.0)
2	mean response time (0.95)	response number (0.65)	total number of fixations (0.95)
3	standard deviation of fixation duration (0.6)	mean saccade amplitude (0.63)	total number of saccades (0.95)
4	total number of fixations (0.53)	standard deviation of fixation duration (0.62)	mean saccade amplitude (0.22)
5	total number of saccades (0.52)	total number of fixations (0.6)	maximum saccade amplitude (0.19)
6	response number (0.27)	total number of saccades (0.55)	mean response time (0.18)
7	standard deviation of pupil diameter (left) (0.17)	maximum fixation duration (0.28)	maximum fixation duration (0.15)
8	maximum fixation duration (0.16)	maximum saccade amplitude (0.15)	total number of blinks (0.1)
9	maximum saccade amplitude (0.5)	total number of blinks (0.09)	standard deviation of pupil diameter (left) (0.09)
10	total number of blinks (0.1)	standard deviation of pupil diameter (left) (0.05)	standard deviation of fixation duration (0.08)

3. Aggregation of Classifiers

Let us recall the most important properties of aggregation operators. Aggregation function p: $[0,\ 1]^n \to [0,\ 1]$ is, in general, defined as an operator fulfilling the following conditions:

$$p(0,0,\ldots,0) = 0, p(1,1,\ldots,1) = 1 \tag{1}$$

and

$$\forall x, y \in [0,\ 1]^n x \le y \Rightarrow P(x) \le p(y) \tag{2}$$

It means that it preserves bounds and monotonicity [31]. Examples are various means or Ordinary Weighted Averaging (OWA) operators [40]. One of the most important and intensively developed aggregation operators is the Choquet integral. To define this integral, we have to recall the properties of fuzzy measure. If X is a set then $Q(X) = 2^X$ is its subsets family. Then a function g fulfilling the conditions

$$g(\varnothing) = 0 \tag{3}$$

$$g(X) = 1 \tag{4}$$

$$g(A) \le g(\beta), \ \ A \subset B, \ \ A, B \in Q(X) \tag{5}$$

$$\lim_{n\to\infty} g(A_n) = g\left(\lim_{n\to\infty} A_n\right) \tag{6}$$

where $\{A_n\}$; $n = 1, 2, \ldots$, denotes an increasing sequence is called fuzzy measure. Note that the Sugeno λ-fuzzy measure is a typical example of fuzzy measure class of functions. Recall that it satisfies

$$g(A \cup B) = g(A) + g(B) + \lambda g(A) g(B) \tag{7}$$

for $\lambda > -1$. Here, A and B are not overlapped. Moreover,

$$g(A_{i+1}) = g(A_i) + g_{i+1} + \lambda g(A_i) \tag{8}$$

where $A_i = \{x_1, \ldots, x_n\}$, $A_{i+1} = \{x_1, \ldots, x_{n+1}\}$. To simplify one writes

$$g_i = g(\{x_i\})\ i = 1, \cdots, n \tag{9}$$

Let $h(x)$ be a function and let $h(x_i)$, $i = 1, \ldots, n$; be ordered in a non-increasing manner. Moreover, let $h(x_{n+1}) = 0$. Then the Choquet integral is

$$CH = \sum_{i=1}^{n} (h(x_i) - h(x_{i+1}) g(A_i)) \tag{10}$$

An interesting generalization for this function is [46,48]

$$C_{MMin}(x) = \sum_{i=1}^{n} M(\min(h(x_i), g(A_i)) - \min(h(x_{i+1}), g(A_i))) \tag{11}$$

or

$$C_{MinM}(x) = \sum_{i=1}^{n} (\min(M(h(x_i), g(A_i)), g(A_i)) - \min(M(h(x_{i+1}), g(A_i)), g(A_i))) \tag{12}$$

Here, M can be any t-norm, see [43,44].

A general model of aggregation processing is presented in Figure 2. The data are classified separately by various classifiers. Next, on a basis of weights, which can be obtained from experts or on a basis of accuracy of individual classifiers, the results are aggregated using a proper aggregation operator.

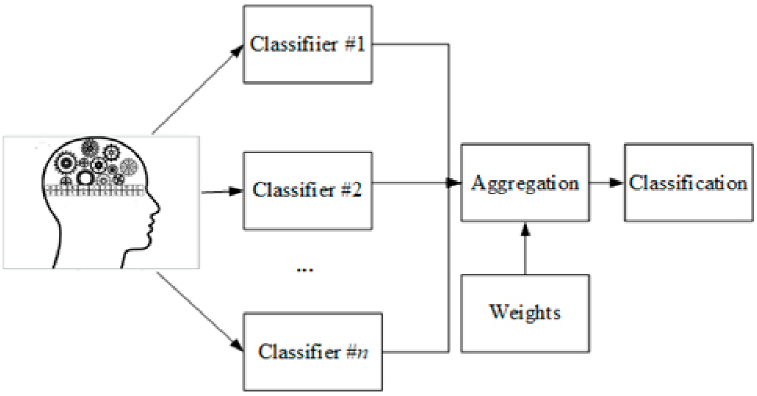


Figure 2. A general aggregation scheme.

4. Experimental Results

4.1. Individual Clssifiers

Several classic machine learning models were tested in the first stage of numerical experiments. The following classifiers were applied: SVMs with various kernels, namely linear, quadratic, and cubic one, Logistic Regression, k-Nearest Neighbors, Decision Tree, Random Forest, Multilayer Perceptron (MLP). Due to the fact that the test sample was balanced, accuracy can be an appropriate classification quality metric. Table 3 shows the mean values of accuracy obtained for various classifiers achieved for both datasets: the dataset containing all 20 features and the dataset containing 10 selected features. It can be noticed from the results, the best classification model allowed to achieve the accuracy reaching the level of 96%. The results show that the classifier accuracy for dataset with selected features are slightly better than the results obtained for all features.

Table 3. Accuracies obtained with separate classifiers.

Model	Accuracy (%) for 10 Selected Features	Accuracy (%) for All Features
SVM(Linear)	94.75	93.11
SVM(Quadratic)	84.47	78.28
SVM(Cubic)	92.36	89.47
Logistic Regression	96.22	94.67
kNN	93.78	89.61
Decision Tree	90.39	90.11
Random Forest	96.22	94.89
MLP	93.53	89.56

Another important aspect worth noting here is the procedure of fuzzy measure density values generation. Several methods of fuzzy measure generation can be used: expert assumption, optimization, and, finally the heuristic one. In our research, we use the heuristic based on cross validation. In order to produce a density measure for a classifier, we run n-fold cross validation on the training set. As the result we obtain n values of accuracy. The mean of cross validation accuracy is considered as the fuzzy measure g_i of the i-th classifier. The fuzzy measures can be interpreted as the degree of trust (or simply weights or level of importance) to a separate classifier's predictions. Figure 3 illustrates the approach.

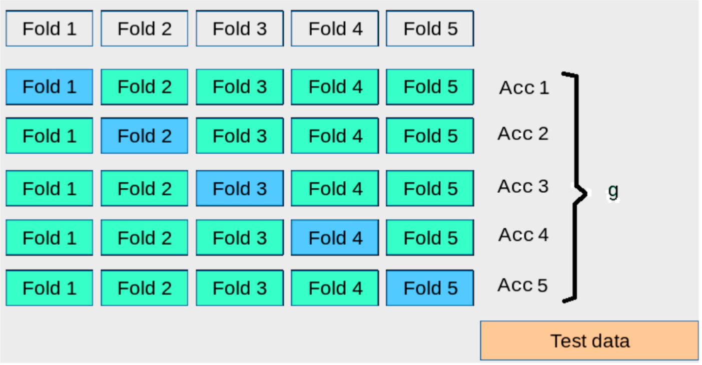


Figure 3. The idea of fuzzy measure generation through the process of cross validation.

4.2. *Aggregation of Classifiers*

Here, we present the best functions serving as aggregation operators for the classifiers listed in the previous subsection, i.e., Cubic SVM, Decision Tree, k-Nearest Neighbor, Linear SVM, Logistic Regression, Multilayer Perceptron, Quadratic SVM, and Random Forest. In the cases where it is needed to feed the aggregation algorithm with weights, they were found on a basis of specific classifiers' accuracy by performing cross validation on training data. For instance, to determine fuzzy measure densities g_i, see Equation (9). The values being the inputs to the aggregation functions are the probabilities of belonging to the three considered classes. Depending on the number of arguments of the specific aggregation function, these values are either provided to a single function or transitive. The latter case is considered when the function has only two arguments. In the validation stage, we considered 200 repetitions, each including tests on 18 validation observations for which we have obtained the probabilities of belonging to the three classes. Let us now discuss the best aggregation operators from over 2000 aggregation operators and so-called pre-aggregation functions (generalized Choquet integrals), see papers [43,45]. The source of the functions were various examples or our own modifications of the functions comprehensively described in [29,31,34,38,50,51] and other books and papers. In the rest of the section, we present the results obtained with particular aggregation operators: both for complete feature set and for selected 10 features. The results are provided in the following format: "selected features result" ("complete feature set result"). The summary of the results is presented on Figure 4.

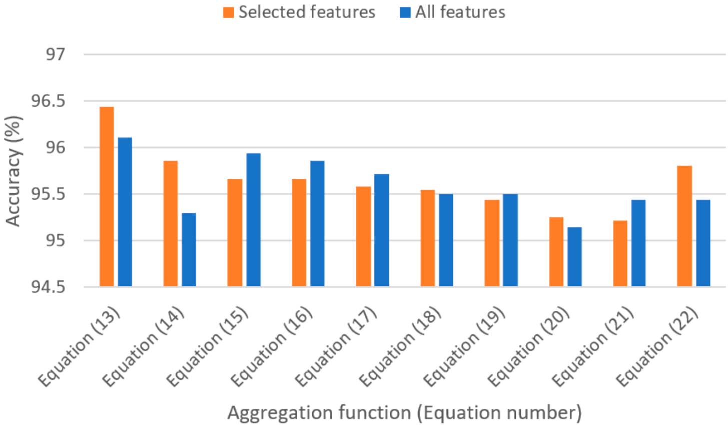


Figure 4. Comparison of accuracies achieved with top aggregation functions on complete feature set and for selected features.

The best result was obtained with a so-called generalized form of Choquet integral [34], i.e.,

$$L(x,y) = \begin{cases} ax + (1-Q)y \ \ for\ x \geq y \\ (1-b)x + by \ \ otherwise \end{cases} \tag{13}$$

where $x \geq 0, y \geq 0, a, b \in [0, 1]$. It gave the accuracy 96.44% (96.11%) for various parameters of a and b, for instance $a = 0.01$, $b = 0.99$. Other selected values of these parameters resulted in correct recognition rates on a slightly lower level. Here, it is worth stressing that the name of the function (12) can be misleading since it is not typical Choquet integral discussed in the previous section, see Equation (10).

The next function producing satisfying results 95.86% (95.3%) is a so-called weighted aggregation function of the form [34]

$$A(x_1, \ldots, x_n) = \frac{\prod_{i=1}^{n}(1+w_i x_i) - \prod_{i=1}^{n}(1-w_i x_i)}{\prod_{i=1}^{n}(1+w_i x_i) + \prod_{i=1}^{n}(1-w_i x_i)} \tag{14}$$

where the values of w_i's are the individual classifiers' accuracies.

The next function, which produces highly satisfying results, is Stolarsky mean [34], [52]

$$M_s(x,y) = \begin{cases} \left(\frac{x^r - y^r}{r(x-y)}\right)^{\frac{1}{r-1}} if\ x \neq y \\ x\ if\ x = y \end{cases} \tag{15}$$

where $r \neq 0$. In this case, the resulting recognition rate is 95.66% (95.94%). For r = 2. The next interesting function is an associative function proposed in [29], namely

$$C(x,y) = \frac{1}{2}W(x,y) + M(x,y) \tag{16}$$

where

$$W(x,y) = \max(x+y-1, 0)$$

and

$$M(x,y) = \frac{x+y}{2}$$

with 95.66% (95.86%) accuracy. A so-called SP-based bivariate symmetric sum [31]

$$f(x,y) = \frac{x+y-xy}{1+x+y-2xy} \tag{17}$$

produced the recognition rate of the level of 95.58% (95.72%). The function of the form

$$f(x,y) = 2^{\log(1+x)\log(1+y)/(\log 2)^2} \tag{18}$$

gave 95.55% (95.5%) recognition rate. The accuracy 95.44% (95.5%) was obtained with an application of a function of the form

$$f(x,y) = \frac{x+y}{2} \tag{19}$$

but if $x \in\ [0.5,\ 0.7)$ the value of x is substituted by 0.5. The same is done with $y \in\ [0.5,\ 0.7)$. Good results are also obtained with a so-called 1-Lipschitzian aggregation function (Bertino copula) [34] (p. 271)

$$f(x,y) = \begin{cases} (\mathrm{Min}(x,y))^2,\ \text{if } x \leq y \\ (\mathrm{Max}(x,y))^2 - |x-y|,\ \text{otherwise} \end{cases} \tag{20}$$

returns 95.25% (95.15%) accuracy. Finally, Sugeno integral [34,50] and max-based bivariate symmetric sum [31], i.e.,

$$f(x,y) = \frac{\max(x,y)}{1+|x-y|} \tag{21}$$

yielded 95.22% (95.44%) recognition rate.

Very good results can also be obtained with the generalization of the Choquet integral of the form (11) and (12). The function M standing under the integral sign was

$$M(x,y) = \left(\ln\left(e^{x^{-\alpha}} + \ln\left(e^{y^{-\alpha}} - e\right)\right)\right)^{-\frac{1}{\alpha}} \tag{22}$$

for $\alpha > 0$. Its value $\alpha = 3.3$ gave the maximal recognition rate at the level of 95.81% (95.44%).

Here, it is worth stressing that also the results at satisfying level were obtained using various fuzzy integrals, most of the pre-aggregation functions or generalized aggregation functions discussed in [38], median or weighted median, scoring or weighted scoring, quadratic mean, and a few versions of ordinary weighted averaging functions (OWA). Interestingly, aggregation operators can improve recognition rate in more noticeable way for the data without extended feature selection.

Figure 4 presents the ranking of the best operators among the tested aggregation functions. The results show that their application affects the quality of classification in a favorable way. The best result, achieved with a generalized form of Choquet integral function, is more than 1.2 percentage point higher for complete feature set and 0.2 percentage point higher for selected features compared to the best individual classifier (Logistic Regression and Random Forest).

5. Discussion

The aim of the study was to improve the result of multiple cognitive workload level classification based on eye activity and user performance. The original classification procedure covering three class classification using classical methods such as SVM, kNN, Decision Tree, Random Forest, MLP, and Logistic Regression was the input to the aggregation functions. In the study, many aggregation and pre-aggregation operators published in the core literature monographs were compared in order to find the best model suitable for classification of cognitive workload level. The results show that using various classification models in combination with an aggregation function allows further improvement of recognition rate by applying the knowledge cumulated in the parameters of the trained models.

The original dataset covering eye-tracking and user performance data was gathered in a study of three parts of the computerized version of DSST test (Digit Symbol Substitution Test). Classification was performed with the interpretable machine learning model in order to regard the most valuable features. Eye-tracking features, in general, have been already proved to be useful in cognitive workload analysis also due to the fact that it is a non-invasive sourced, natural type of response obtained without additional activity or training. What is more, the classification was performed as subject-independent in order to distinguish classes regardless of such conditions as the age of an examined person, his/her habits, or testing period. The best original classification results achieved 96%. It is worth noting that the tests were performed on a homogeneous group of healthy people with similar age and educational level.

The study presented in the paper proved that applying aggregation methods enables to increase the classification result by more than 1 percentage point. Detailed results show that there were several aggregation functions that enabled achieving the highest results (presented in the paper are the top ten functions as Equations (13)–(22)).

Classification results, both individual and with aggregation, prove that the time and difficulty level of performed tasks have a systematic influence on user performance, pupillary and eye movements. The results show that there is a relation between the participants' engagement combined with cognitive state and eye activity. The most important features in the study are these related to the user performance and the intensity of eye movement. It indicates that fixation and saccade-related features (mean saccade amplitude, standard deviation of fixation duration, total number of fixations and saccades) as well as response-related features (mean response time, response number) reflect the degree of attention during the tasks performance. However, further results are needed to investigate additional factors such as types of tasks, participant profiles or their initial mental state. What is more, it is worth to consider the mental abilities of each single participant. Such information might help to adjust the cognitive workload to a particular participant. This might be measured with dedicated models or surveys (e.g., NASA-TLX scale, the Rasch and strain–stress model), although such tools are based on subjective assessment.

A broad set of pre-aggregation and aggregation operators was analyzed in the study in order to find the ones that fit the best to the analyzed problem. The detailed results show that the classification accuracy was improved.

In the case study, two approaches were applied. The first one was based on classification considering original 20 features whereas the second one covered 10 features chosen in statistical analysis. The individual classification results for both approaches differ slightly, although the results for smaller number of features occurred to be better. Results for both approaches were further processed in order to apply pre-aggregation and aggregation operators. The best results for both approaches were achieved for the generalized Choquet integral. This operator enabled to improve the classification results by as much as 1.2 percentage point for all features-based approach compared to the best classification model. The same operator proved to be efficient also in case of a smaller feature number approach, although the improvement was not as high. It was Random Forest that occurred to be the best among the classical classifiers for both approaches. Additionally, Logistic Regression gave similar results for the second approach. These results confirm usefulness of the generalized Choquet integral found in research over classification performance. The results prove that the application of pre-aggregation and aggregation operators is useful especially in case of applying the basic feature selection. Aggregation functions might give better improvement in case of weaker initial individual classification results.

Future work is planned to include the experiments on a broader dataset, collected from a higher number of participants. The authors also consider analysis of a higher number of cognitive workload levels. As further development of the topic, it is planned to include self-report tools of detecting mental illness such as depression or anxiety symptoms in our future work.

Author Contributions: Conceptualization, M.K., P.K., M.P.-W. and M.T.; methodology, P.K., M.K. and M.T.; software, M.K. and P.K.; validation, M.K., P.K., M.P.-W. and M.T.; data curation, M.K.; writing—original draft preparation, M.K., P.K., M.P.-W. and M.T.; writing—review and editing, M.K., P.K., M.P.-W. and M.T.; visualization, M.K. and M.T.; supervision, M.P.-W.; project administration, M.P.-W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Ethics Committee of Lublin University of Technology (Approval ID 2/2018 from 19 October 2018).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Qi, P.; Ru, H.; Gao, L.; Zhang, X.; Zhou, T.; Tian, Y.; Sun, Y. Neural mechanisms of mental fatigue revisited: New insights from the brain connectome. *Engineering* **2019**, *5*, 276–286. [CrossRef]
2. Chen, L.L.; Zhao, Y.; Ye, P.F.; Zhang, J.; Zou, J.Z. Detecting driving stress in physiological signals based on multimodal feature analysis and kernel classifiers. *Expert Syst. Appl.* **2017**, *85*, 279–291. [CrossRef]
3. Wang, Z.; Hope, R.M.; Wang, Z.; Ji, Q.; Gray, W.D. Cross-subject workload classification with a hierarchical Bayes model. *NeuroImage* **2012**, *59*, 64–69. [CrossRef]
4. Walter, C.; Wolter, P.; Rosenstiel, W.; Bogdan, M.; Spüler, M. Towards cross-subject workload prediction. In Proceedings of the 6th International Brain-Computer Interface Conference, Graz, Austria, 16–21 September 2014.
5. Thodoroff, P.; Pineau, J.; Lim, A. Learning robust features using deep learning for automatic seizure detection. In Proceedings of the Machine learning for healthcare conference, Los Angeles, CA, USA, 19–20 August 2016; pp. 178–190.
6. McKendrick, R.; Feest, B.; Harwood, A.; Falcone, B. Theories and methods for labeling cognitive workload: Classification and transfer learning. *Front. Hum. Neurosci.* **2019**, *13*, 295. [CrossRef]
7. Fridman, L.; Reimer, B.; Mehler, B.; Freeman, W.T. Cognitive load estimation in the wild. In Proceedings of the 2018 Chi Conference on Human Factors in Computing Systems, Montreal, QC, Canada, 21–26 April 2018; pp. 1–9.
8. Appel, T.; Scharinger, C.; Gerjets, P.; Kasneci, E. Cross-subject workload classification using pupil-related measures. In Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications, Warsaw, Poland, 14–17 June 2018; pp. 1–8.
9. Almogbel, M.A.; Dang, A.H.; Kameyama, W. EEG-signals based cognitive workload detection of vehicle driver using deep learning. In Proceedings of the 2018 20th International Conference on Advanced Communication Technology (ICACT), Chuncheon, Korea, 11–14 February 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 256–259.
10. Hefron, R.; Borghetti, B.; Schubert Kabban, C.; Christensen, J.; Estepp, J. Cross-participant EEG-based assessment of cognitive workload using multi-path convolutional recurrent neural networks. *Sensors* **2018**, *18*, 1339. [CrossRef] [PubMed]
11. Lobo, J.L.; Ser, J.D.; De Simone, F.; Presta, R.; Collina, S.; Moravek, Z. Cognitive workload classification using eye-tracking and EEG data. In Proceedings of the International Conference on Human-Computer Interaction in Aerospace, Paris, France, 14–16 September 2016; pp. 1–8.
12. Almogbel, M.A.; Dang, A.H.; Kameyama, W. Cognitive workload detection from raw EEG-signals of vehicle driver using deep learning. In Proceedings of the 2019 21st International Conference on Advanced Communication Technology (ICACT), PyeongChang, Korea, 17–20 February 2019; pp. 1–6.
13. Zarjam, P.; Epps, J.; Chen, F.; Lovell, N.H. Estimating cognitive workload using wavelet entropy-based features during an arithmetic task. *Comput. Biol. Med.* **2013**, *43*, 2186–2195. [CrossRef]
14. Chen, J.; Wang, H.; Wang, Q.; Hua, C. Exploring the fatigue affecting electroencephalography based functional brain networks during real driving in young males. *Neuropsychologia* **2019**, *129*, 200–211. [CrossRef] [PubMed]
15. Yamada, Y.; Kobayashi, M. Detecting mental fatigue from eye-tracking data gathered while watching video: Evaluation in younger and older adults. *Artif. Intell. Med.* **2018**, *91*, 39–48. [CrossRef]
16. Khushaba, R.N.; Kodagoda, S.; Lal, S.; Dissanayake, G. Driver drowsiness classification using fuzzy wavelet-packet-based feature-extraction algorithm. *IEEE Trans. Biomed. Eng.* **2010**, *58*, 121–131. [CrossRef]
17. Maiorana, E. Deep learning for EEG-based biometric recognition. *Neurocomputing* **2020**, *410*, 374–386. [CrossRef]
18. Hajinoroozi, M.; Mao, Z.; Jung, T.P.; Lin, C.T.; Huang, Y. EEG-based prediction of driver's cognitive performance by deep convolutional neural network. *Signal Process. Image Commun.* **2016**, *47*, 549–555. [CrossRef]
19. Wobrock, D.; Frey, J.; Graeff, D.; De La Rivière, J.B.; Castet, J.; Lotte, F. Continuous mental effort evaluation during 3d object manipulation tasks based on brain and physiological signals. In Proceedings of the IFIP Conference on Human-Computer Interaction, Bamberg, Germany, 14–18 September 2015; Springer: Cham, Switzerland; Berlin/Heidelberg, Germany, 2015; pp. 472–487.
20. Bozkir, E.; Geisler, D.; Kasneci, E. Person independent, privacy preserving, and real time assessment of cognitive load using eye tracking in a virtual reality setup. In Proceedings of the 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), Osaka, Japan, 23–27 March 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1834–1837.
21. Zhao, Z.; Wang, C.; Niu, Y.; Shen, L.; Ma, Z.; Wu, L. Adjustable Autonomy for Human-UAVs Collaborative Searching Using Fuzzy Cognitive Maps. In Proceedings of the 2019 2nd China Symposium on Cognitive Computing and Hybrid Intelligence (CCHI), Xi'an, China, 21–22 September 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 230–234.
22. Naqvi, R.A.; Arsalan, M.; Park, K.R. Fuzzy system-based target selection for a NIR camera-based gaze tracker. *Sensors* **2017**, *17*, 862. [CrossRef] [PubMed]
23. Yusuf, A.B.; Kor, A.L.; Tawfik, H. Development of a Simulation Experiment to Investigate In-Flight Startle using Fuzzy Cognitive Maps and Pupillometry. In Proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, 14–19 July 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–10.
24. Fatimah, B.; Pramanick, D.; Shivashankaran, P. Automatic detection of mental arithmetic task and its difficulty level using EEG signals. In Proceedings of the 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kharagpur, India, 1–3 July 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–6.
25. Han, S.Y.; Kwak, N.S.; Oh, T.; Lee, S.W. Classification of pilots' mental states using a multimodal deep learning network. *Biocybern. Biomed. Eng.* **2020**, *40*, 324–336. [CrossRef]

26. Becerra-Sánchez, P.; Reyes-Munoz, A.; Guerrero-Ibañez, A. Feature selection model based on EEG signals for assessing the cognitive workload in drivers. *Sensors* **2020**, *20*, 5881. [CrossRef] [PubMed]
27. Dell'Agnola, F.; Momeni, N.; Arza, A.; Atienza, D. Cognitive workload monitoring in virtual reality based rescue missions with drones. In Proceedings of the International Conference on Human-Computer Interaction, Copenhagen, Denmark, 19–24 July 2020; Springer: Cham, Switzerland; Berlin/Heidelberg, Germany, 2015; pp. 397–409.
28. Chakladar, D.D.; Dey, S.; Roy, P.P.; Iwamura, M. EEG-Based Cognitive State Assessment Using Deep Ensemble Model and Filter Bank Common Spatial Pattern. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 4107–4114.
29. Alsina, C.; Schweizer, B.; Frank, M.J. *Associative Functions: Triangular Norms and Copulas*; World Scientific: Singapore, 2006.
30. Karczmarek, P.; Kiersztyn, A.; Pedrycz, W. Generalizations of Aggregation Functions for Face Recognition. In Proceedings of the International Conference on Artificial Intelligence and Soft Computing, Zakopane, Poland, 16–20 June 2019; pp. 182–192.
31. Beliakov, G.; Pradera, A.; Calvo, T. *Aggregation Functions: A Guide for Practitioners*; Springer: Berlin/Heidelberg, Germany, 2007; Volume 221.
32. Calvo, T.; Mayor, G.; Mesiar, R. (Eds.) *Aggregation Operators: New Trends and Applications*; Physica: Amsterdam, The Netherlands, 2012; Volume 97.
33. Gągolewski, M. *Data Fusion: Theory, Methods, and Applications*; Institute of Computer Science, Polish Academy of Sciences: Warszawa, Poland, 2015.
34. Grabisch, M.; Marichal, J.L.; Mesiar, R.; Pap, E. *Aggregation Functions (No. 127)*; Cambridge University Press: Cambridge, UK, 2009.
35. Mesiar, R.; Kolesárová, A.; Calvo, T.; Komorníková, M. A review of aggregation functions. Fuzzy sets and their extensions: Representation, aggregation and models. *Stud. Fuzziness Soft Comput.* **2008**, *220*, 121–144.
36. Grabisch, M.; Marichal, J.L.; Mesiar, R.; Pap, E. Aggregation functions: Means. *Inf. Sci.* **2011**, *181*, 1–22. [CrossRef]
37. Grabisch, M.; Marichal, J.L.; Mesiar, R.; Pap, E. Aggregation functions: Construction methods, conjunctive, disjunctive and mixed classes. *Inf. Sci.* **2011**, *181*, 23–43. [CrossRef]
38. Klement, E.P.; Mesiar, R.; Pap, E. *Triangular Norms*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2013; Volume 8.
39. Klement, E.P.; Mesiar, R. (Eds.) *Logical, Algebraic, Analytic and Probabilistic Aspects of Triangular Norms*; Elsevier: Amsterdam, The Netherlands, 2005.
40. Yager, R.R.; Kacprzyk, J. (Eds.) *The Ordered Weighted Averaging Operators: Theory and Applications*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2012.
41. Choquet, G. Theory of capacities. *Annales de l'institut Fourier* **1954**, *5*, 131–295. [CrossRef]
42. Grabisch, M. The application of fuzzy integrals in multicriteria decision making. *Eur. J. Oper. Res.* **1996**, *89*, 445–456. [CrossRef]
43. Bustince, H.; Sanz, J.A.; Lucca, G.; Dimuro, G.P.; Bedregal, B.; Mesiar, R. Pre-aggregation functions: Definition, properties and construction methods. In Proceedings of the 2016 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE) 2016, Hyderabad, India, 7–10 July 2013; pp. 294–300.
44. Karczmarek, P.; Pedrycz, W.; Kiersztyn, A.; Dolecki, M. A comprehensive experimental comparison of the aggregation techniques for face recognition. *Iran. J. Fuzzy Syst.* **2019**, *16*, 1–19.
45. Lucca, G.; Sanz, J.A.; Dimuro, G.P.; Bedregal, B.; Mesiar, R.; Kolesárová, A.; Bustince, H. The notion of pre-aggregation function. In Proceedings of the International Conference on Modeling Decisions for Artificial Intelligence 2015, Skövde, Sweden, 21–23 September 2015; Springer: Cham, Switzerland; Berlin/Heidelberg, Germany, 2015; pp. 33–41.
46. Karczmarek, P.; Kiersztyn, A.; Pedrycz, W. Generalized choquet integral for face recognition. *Int. J. Fuzzy Syst.* **2018**, *20*, 1047–1055. [CrossRef]
47. Dimuro, G.P.; Fernández, J.; Bedregal, B.; Mesiar, R.; Sanz, J.A.; Lucca, G.; Bustince, H. The state-of-art of the generalizations of the Choquet integral: From aggregation and pre-aggregation to ordered directionally monotone functions. *Inf. Fusion* **2020**, *57*, 27–43. [CrossRef]
48. Karczmarek, P. *Selected Problems of Face Recognition and Decision-Making Theory*; Wydawnictwo Politechniki Lubelskiej: Lublin, Poland, 2018.
49. Boake, C. From the Binet–Simon to the Wechsler–Bellevue: Tracing the history of intelligence testing. *J. Clin. Exp. Neuropsychol.* **2002**, *24*, 383–405. [CrossRef] [PubMed]
50. Pedrycz, W.; Gomide, F. *An Introduction to Fuzzy Sets: Analysis and Design*; MIT Press: Cambridge, MA, USA, 1988.
51. Torra, V.; Narukawa, Y. *Modeling Decisions: Information Fusion and Aggregation Operators*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2007.
52. Stolarsky, K.B. Generalizations of the logarithmic mean. *Math. Mag.* **1975**, *48*, 87–92. [CrossRef]

6.4 Automated Classification of Cognitive Workload Levels Based on Psychophysiological and Behavioural Variables of Ex-Gaussian Distributional Features

1. Details

This article was written by Monika Kaczorowska, Małgorzata Plechawska-Wójcik, Mikhail Tokovarov and Paweł Krukow and published in Brain sciences in 2022, vol. 12, nr 5. This article is worth 100 points according to the list of the Polish Ministry of Science and Higher Education.

2. Abstract

This study is focused on applying ex-Gaussian parameters of eye-tracking and cognitive measures in the classification process of cognitive workload level. A computerized version of the digit symbol substitution test has been developed in order to perform the case study. The dataset applied in the study is a collection of variables related to eye-tracking: saccades, fixations and blinks, as well as test-related variables including response time and correct response number. The application of ex-Gaussian modelling to all collected data was beneficial in the context of detection of dissimilarity in groups. An independent classification approach has been applied in this study. Several classical classification methods have been invoked in the process. The overall classification accuracy reached almost 96%. Furthermore, the interpretable machine learning model based on logistic regression was adapted in order to calculate the ranking of the most valuable features, which allowed us to examine their importance.

brain sciences

MDPI

Article

Automated Classification of Cognitive Workload Levels Based on Psychophysiological and Behavioural Variables of Ex-Gaussian Distributional Features

Monika Kaczorowska [1], Małgorzata Plechawska-Wójcik [1], Mikhail Tokovarov [1] and Paweł Krukow [2,*]

[1] Department of Computer Science, Lublin University of Technology, 20-618 Lublin, Poland; m.kaczorowska@pollub.pl (M.K.); m.plechawska@pollub.pl (M.P.-W.); m.tokovarov@pollub.pl (M.T.)

[2] Department of Clinical Neuropsychiatry, Medical University of Lublin, 20-439 Lublin, Poland

[*] Correspondence: pawel.krukow@umlub.pl; Tel.: +48-665-226-144

Abstract: The study is focused on applying ex-Gaussian parameters of eye-tracking and cognitive measures in the classification process of cognitive workload level. A computerised version of the digit symbol substitution test has been developed in order to perform the case study. The dataset applied in the study is a collection of variables related to eye-tracking: saccades, fixations and blinks, as well as test-related variables including response time and correct response number. The application of ex-Gaussian modelling to all collected data was beneficial in the context of detection of dissimilarity in groups. An independent classification approach has been applied in the study. Several classical classification methods have been invoked in the process. The overall classification accuracy reached almost 96%. Furthermore, the interpretable machine learning model based on logistic regression was adapted in order to calculate the ranking of the most valuable features, which allowed us to examine their importance.

Keywords: ex-Gaussian modelling; classical machine learning; cognitive workload

check for updates

Citation: Kaczorowska, M.; Plechawska-Wójcik, M.; Tokovarov, M.; Krukow, P. Automated Classification of Cognitive Workload Levels Based on Psychophysiological and Behavioural Variables of Ex-Gaussian Distributional Features. *Brain Sci.* **2022**, *12*, 542. https://doi.org/10.3390/brainsci12050542

Academic Editor: Mera S Barr

Received: 14 February 2022
Accepted: 21 April 2022
Published: 23 April 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Cognitive load classification is the subject of numerous scientific studies [1–3]. Its goal is to verify whether there are objective indicators of several cognitive workload levels enabling its recognition by various computational methods. The authors carried out both subject-dependent [4] and subject-independent [5] classification inquiries; the latter becoming progressively more popular in the research. The subject-independent classification approach is where samples taken from a single subject are fully included in one dataset (train or test) in order to ensure reliability of results. In the literature, one can also find utilization of both approaches applied simultaneously [6]. The most common classification of the level of cognitive load is the binary approach (presence or absence of overload) [7], but there are also attempts to classify several levels of cognitive load [8]. The literature review shows that the cognitive load is most often tested on the basis of eye-tracking signals [8] and EEG [9]. As for classification methods, both classic classifiers and deep neural networks [10,11] are administrated. In terms of classical classifiers one can mention such models as support vector machine (SVM) [1,12], linear discriminant (LDA) [3] and k-nearest neighbours (kNN) [13]. Depending on the type of classification, the authors usually present their results between 70–95% accuracy. Yamada and Kobayashi [14] conducted a study including 12 participants and undertook a two-class classification of cognitive load with the use of the SVM classifier, achieving a result of over 90% accuracy. The model created classified the effort or its lack, regardless of the age of the examined person. In other research [15], the authors presented a seven-class classification, which concerned the sorting of cognitive load on the basis of the eye-tracking signal when performing arithmetic tasks, from the easiest to the most difficult. The authors obtained an accuracy reaching from 0.4 to 0.98 using artificial neural networks. There are other studies presenting the results of binary [16–18] and three-class classification [6,19] of cognitive workload. As suggested earlier, the classification studies on cognitive workload have two main purposes: to find the objective indicators of cognitive overburden (e.g., behavioural, cognitive and physiological) and to verify computational methods enabling the most accurate automatic recognition and differentiation of overload levels.

The vast majority of cognitive experimental data sets consisting of a large amount of numeric results are still analysed with the application of parametric analytical methods, first of all in the form of means and standard deviations. Despite this, researchers focused on measuring reaction times (RTs) have noticed that the distribution of such outcomes, even when collected from healthy participants, is often skewed and contrary to expectations; the empirical distribution of RT data does not always fit the Gaussian model on which parametric statistics are based [20]. Typically, the skewness of the distribution of the RT series is positive; that is, it has a set of the most typical and at the same time relatively short, most frequent RTs, and an extended right-side convolution tail, containing the rarest, but at the same time longest RTs, representing the most unusual, prolonged RT outliers of the whole distribution [21]. Such a form of data distribution suggests that in a series of RTs parametric symmetrical standard deviation (SD) does not fully cover the real range of outliers, and instead, averaging all data to the form of arithmetic means eliminates some portion of uncommon observations and ultimately makes it impossible to assess the true extent of intra-individual variability (IIV). Hence, to circumvent some limitations of the parametric approach to experimental RT-based data, the so-called ex-Gaussian methodology has been developed, allowing analysis of positively skewed distribution so as not to eliminate outliers, and at the same time, not distorting the range of typical results. Ex-Gaussian distribution modelling provides quantitative characteristics in the form of three independent parameters: mu (μ), representing the mean of the normal component and reflecting average performance or the most frequent results; sigma (σ), defining the symmetrical standard deviation of the normal component; and tau (τ), covering the exponential part of the distribution with the most prolonged and most often rarest RTs [22]. In cases when the experimental procedure consists of displaying repetitive stimuli and/or relatively unified reactions are expected, it is considered that the τ parameter covering the scope of the most prolonged RTs might be understood as an indicator of “attentional lapses” or “off-task mind wandering”. Attentional lapses are usually due to transient failures in performance controlling mechanisms occurring, for example, due to increasing cognitive fatigue [23–25]. RT-based and other studies focused on cognitive performance and its disturbances confirm that an increased range of attentional lapses and off-task mind wandering is associated with lower levels of effortful control [26], states of reduced alertness and generally diminished productivity [27,28], and additionally with mental health risk factors, such as anxiety and negative affect [29,30].

Although measures of intra-individual variability, including ex-Gaussian modelling, have so far been used mainly to analyse features of RT distribution, it does not mean that these methods cannot be successfully implemented for other types of empirical data. For example, indicators of intra-subject variance were used to analyse neuroimaging results, especially regarding patterns of variability in neuronal functional connectivity matrices [31,32], analogous computation has been exploited in studies on heart rate variability [32–34] and ex-Gaussian modelling enabled the highlighting of intergroup differences in the extremely high measures of IgG allergic reaction markers in patients with depression, irritable bowel syndrome and healthy controls [35].

It can be argued without a doubt that, just as the application of ex-Gaussian modelling to cognitive data in the form of RT series is relatively common, the use of this quantitative analysis method for eye-tracking output is still scarce. To our knowledge, one of the first studies in which the distribution of variables such as fixation length and intersaccadic intervals were successfully matched to ex-Gaussian distribution was carried out by Otero-Millian et al. [36]. In the following years, the ex-Gaussian modelling of eye-tracking results was described only a few more times, while probably the fullest theoretical and empirical justification for the application of distributional analyses to parameters such as fixation length was presented by Guy, Lancry-Dayan and Pertzov [37]. These authors documented not only the parametric mean of fixation duration (FD), but also that its empirical distribution is sensitive to experimental manipulations of the task input; they also alluded to the results of previous studies which confirmed the existence of significant relationships between task features such as semantic clarity, stimuli familiarity and cognitive individual differences, e.g., regarding working memory, and FD ex-Gaussian characteristics, especially μ and τ. First of all, in their original study Guy and co-workers evidenced that under three different experimental conditions FD distributions fitted the ex-Gaussian curves even better than the Gaussian one; μ and τ cover significantly different aspects of eye movements and are not redundant or overlapping variables. Additionally, when the same individuals were subjected to the same eye-tracking experiments with a 7-day interval, ex-Gaussian parameters exhibited very high reliability (the correlation between the first and second assessments was at least 0.80). In the end, Guy et al. [17] argued that the τ parameter is associated with repetitive exposures to the same images, while μ, as previously suggested, is to a greater extent related to stimuli familiarity and individual differences in problem-solving efficiency. The increase in τ calculated from FD might represent a psychophysiological marker of attentional lapses because as regards the repetitive presentation of stimuli, i.e., an experimental situation, when a testee already knows given stimuli and the duration of individual fixations enlarges, it is rather unlikely that such extremely prolonged fixation represents the process of stimulus decoding since it has already been visually decoded. Although in the discussed research ex-Gaussian modelling was applied only to one variable, which, as RTs, has a temporal character (duration), we postulate that since the distribution has an axis representing the frequency of measured observations, this approach may also be used to analyse non-temporal eye movement features, such as the magnitude or amplitude of microsaccades and saccades. Then, μ might cover the typical, most recurring observations, while τ will refer to the rarest and to some extent extreme attributes of eye-movements. Taken together, we assume that application of ex-Gaussian modelling to data collected from eye-tracking during exposure to tasks eliciting cognitive workload seems to be fully justified. Additionally, we expect that both μ and τ measures will prove to be significant predictors allowing objective classification of cognitive workload levels with high accuracy.

At this point, we would like to assert that according to our knowledge, although the eye-tracking data were analysed with an application of ex-Gaussian modelling, it was not implemented regarding data taken from the cognitive workload experiment. We presume, that increasing cognitive overburden is associated with a growing number of atypical RTs and atypical physiological events, and therefore, utilization of ex-Gaussian parameters may enhance the possibilities of its automated classification.

In the current study, we implemented ex-Gaussian modelling to analyse two types of data: cognitive variables in the form of RTs and correct or invalid reactions together with psychophysiological variables collected from the eye-tracker. Eye-tracking values cover data related to saccades, fixations and blinks. Both of these groups of data were gathered during an experimental procedure whose aim was to elicit the state of cognitive overburden. In comparison with the aforementioned Guy et al. study [17], we decided to corroborate our premises according to which ex-Gaussian characteristics might reflect not only time-related variables' distributions (e.g., fixation duration) but also another type of psychophysiological measures collected during the cognitive workload experiment. This might be acknowledged as potentially original input in our research.

Considering the above, our study has two main goals:

- To verify whether the cognitive and physiological data collected during cognitive-workload-related experiments fit the ex-Gaussian distribution;
- To determine the possibilities of machine-learning-based classifiers regarding automatic recognition of cognitive workload using ex-Gaussian parameters of eye-tracking and cognitive measures.

Taking into account the above goals, this manuscript is organized as follows: The Materials and Methods section describes the research procedures covering the computer application, equipment, experiment details, data processing, classification methods and statistical procedures. The Results section contains distributional analyses regarding the fitness of the obtained data to ex-Gaussian curves and classifications outcomes. The last section (Discussion) describes and explains the results with reference to the study goals.

2. Materials and Methods

2.1. Research Procedure

Obtained experimental input collected in the current study is fully original and does not coincide with our previous research. The experiment itself consisted of administrating a computerised version of the digit symbol substitution test (DSST) [38], which measures the speed of cognitive processing. While performing DSST, a subject's task was to connect the abstract symbols to the corresponding numbers as quickly as possible. In order to match the number with the symbol, the participant had to indicate with the mouse an appropriate number on a displayed keyboard. The experiment was divided into three parts. In each part, the participant's task was to match the number with the symbol, but the parts differed regarding the number of symbols and duration of the task itself. The first part lasted 90 s and had 4 symbols to choose from, the second and third parts had 9 symbols, but the second part lasted 90 s and the third part lasted 180 s. Thus, it can be assumed that the level of cognitive load in each successive part was higher than in the previous one, giving in sum three levels of cognitive workload. The application was written in Java 8.0.

Figure 1 shows the procedure of the experiment. Each part consisted of the same steps. The successive parts differed in the difficulty of the task being performed.

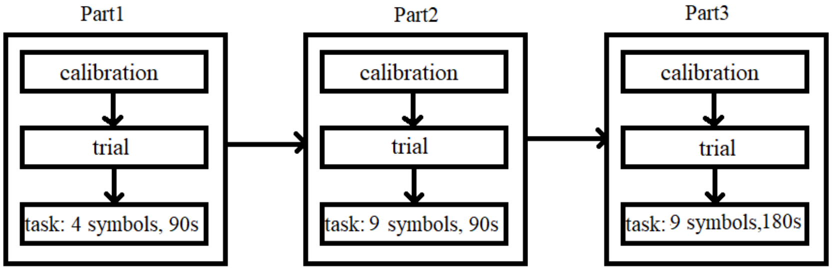


Figure 1. The procedure of the experiment.

2.2. Data Acquisition

The experiment was carried out in the laboratory using the proprietary application and the stationary eye-tracker Tobii Pro TX300 (Tobii AB, Stockholm, Sweden). The eye activity-related data were recorded with 300 Hz frequency. The Tobii Studio 3.2, a software compatible with the eye-tracker, was used in the described experiment. In general, the experiment lasted about 15 min. A nine-point calibration was performed at the start of each measurement. Then the participant proceeded to implement each of the parts. Before each part, the trial version was presented so that the participant knew what to do in a given part. Using the Tobii Studio 3.2 software, the eye activities were exported and saved to an individual file. Additionally, the application generated a file with information about the answers given by the participant. The accepted quality of eye-tracking recording was set to 90% of eye activity.

Thirty healthy participants (students) were involved in the study. The group included subjects aged 20 to 24 (M = 20.45; SD = 1.62); 23 males and 7 females. Additionally, participants had no history of psychiatric nor neurologic diagnoses and reported not to be undergoing medical treatment or taking medication. All participants had normal/corrected-to-normal vision.

2.3. Data Processing

The dataset applied in the study covered features related to eye-activity and cognitive results of the DSST performance. The following eye-activity-related measures, obtained from the eye-tracking equipment, were used in the study:

- Saccades [39], understood as eyes movements bringing the essential visual information onto the most sensitive part of the retina. This process is performed to retrieve information easily [40].
- Fixations [39], described as the time period when the visual information is proceeded. During that time eyes stay in a relatively stable position.
- Blinks, identified by Tobii Studio software as zero data saccades [41].

Among the cognitive measures received from the DSST application we applied the following:

- Response time defined as the time needed to perform a single matching in the application.
- Good response numbers understood as the number of correct answers given in a certain time period.

The abovementioned variables were used in the data processing procedure, which covered such steps as data synchronisation, feature extraction, feature selection and classification.

The synchronisation procedure was related to the process of merging of data received from eye-tracker equipment and from the DSST application. Synchronisation was made on the basis of timestamps saved in both datasets. As each participant took part in three parts of the assessment, each with a different level of cognitive workload, 90 observations were included in the final dataset.

Synchronised outputs were subjected to the basic data cleansing procedure. One highest and one lowest value were deleted from each data series obtained from a particular subject.

The feature extraction procedure was performed using ex-Gaussian statistics. We decided to use this method as it offers a good prognosis [35] for dissimilarity detection in groups. Ex-Gaussian parameters allow us to consider the exponential specificity of the data. The ex-Gaussian distribution enables us to distinguish three independent parameters:

- Mu (μ)—corresponding to the mean of the normal component;
- Sigma (σ)—representing the symmetric standard deviation of the normal component;
- Tau (τ)—reflecting the exponential part of the distribution.

Mu and sigma in ex-Gaussian modelling correspond to classical mean and standard deviation. Mu results were used in the final dataset in order to consider averaged, most frequently occurring results. The tau parameter was included as it indicates the extremes in results dispersion, or outliers usually eliminated in evaluations based on normal distributions. Ex-Gaussian modelling was accomplished using the MATLAB toolbox "DISTRIB" and the recommendations of Lacouture and Cousineau [22].

Ex-Gaussian parameters were calculated for the following eye-related measures: saccades (amplitude of saccade, number of saccades and saccade duration), fixations (number of fixations and fixation duration) and blinks (number of blinks) as well as for DSST-related measures: number of correct answers and single trial response time. Among the listed features, the number of saccades, fixations, blinks and correct responses were extracted for the specific time intervals (10 s period). The taking into account of cognitive and psychophysiological data extracted for shorter intervals was dictated by the conclusions of our earlier research showing that information processing is specifically time-organised and shows variable dynamics and temporal changes in its efficiency [42–44]. Therefore, we decided that isolating the potential dynamic dimension of task performance may also in this case strengthen the effectiveness of classification of cognitive workload at various levels.

In the feature selection procedure the nonparametric Friedman test ($\alpha = 0.05$) was applied to check the significance of the features. Results showed that most of the sigma parameters (for such features as number of fixations, fixation duration, number of saccades, saccade duration, number of correct answers and single trial response time) were non-significant. Taking into account these results and the fact that the sigma parameters indicate only how far the data are spread from the mean, we decided to discard all the sigma parameters from the further analysis in order to reduce the data dimensionality. Ultimately the resulting dataset contained the following 16 features:

- Saccade-related features: mu of saccade amplitude, tau of saccade amplitude, mu of saccade duration, tau of saccade duration, mu of saccade number in 10 s, tau of saccade number in 10 s;
- Fixation-related features: mu of fixation duration, tau of fixation duration, mu of fixation number in 10 s, tau of fixation number in 10 s;
- Blink-related features: mu of blink number in 10 s, tau of blink number in 10 s;
- DSST-related measures: mu of correct answers number in 10 s, tau of correct answers number in 10 s, mu of single trial response time, tau of single trial response time.

Additionally, correlation between particular features were checked using Spearman's correlation coefficient with a significance level of $\alpha = 0.05$. There were no significant strong correlations found in the dataset.

The classification procedure was performed in order to distinguish between three classes: low, medium and high cognitive workload. These three classes correspond to three parts of the DSST-based experiment. The dataset was randomly divided into the training and testing parts in the ratio 80:20. A subject-independent approach was applied in the procedure and data from a single participant were assigned to only one dataset (train or test) in order to ensure full independence of both datasets. Several classification algorithms such as decision tree, SVM, random forest and logistic regression classifier were applied in the study. Furthermore, feature weights were extracted by using the logistic regression model in order to assess the importance of particular features in the process of distinguishing cognitive workload levels. The entire train–test division and classification procedure was independently repeated 200 times in order to achieve reliable classification results.

3. Results

3.1. Distributional Analyses

The first step of the outcome analysis consisted of checking whether the data fit the ex-Gaussian distribution. As depicted in Figure 2, the majority of the studied variables' empirical distribution matched to ex-Gaussian characteristics. In detail, Figure 2 presents distributions of measures such as amplitude of saccade, number of saccades, saccade duration, number of fixations, fixation duration, number of blinks, number of correct answers and single trial response time. Additionally, charts show the right-side exponential tail related to tau parameters.

The level of skewness was positive and constantly grew for measures at all cognitive workload levels, e.g., saccade duration (values 0.2231, 0.4379 and 0.4522, respectively) and response time (values 3.7235, 3.8627 and 4.0802, respectively).

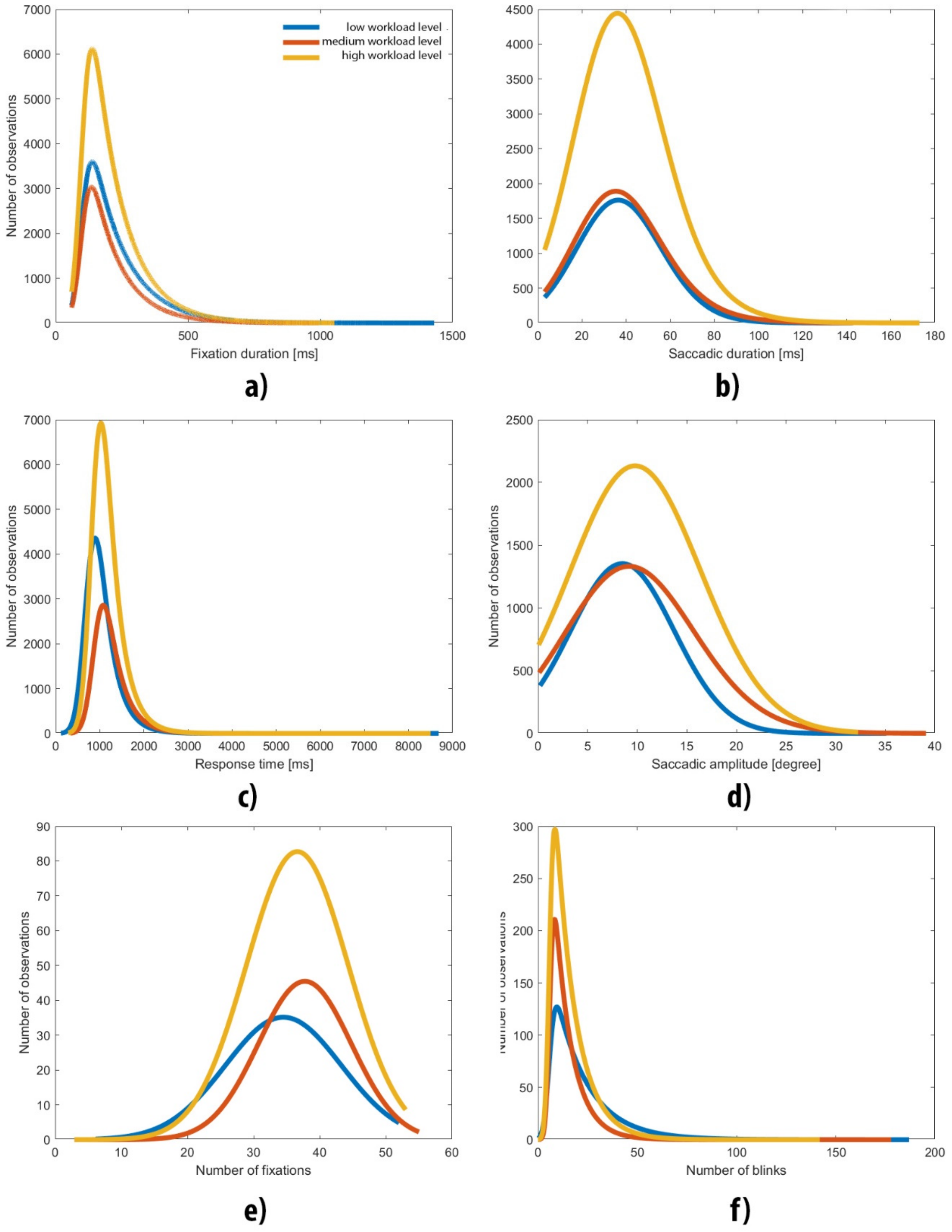


Figure 2. Distribution of features. Level 1 stands for low cognitive workload (**a**)—fixation duration, (**b**)—saccade duration, (**c**)—response time, (**d**)—saccade amplitude, (**e**)—number of fixations, (**f**)—number of blinks).

3.2. Classification Results

A tree-class cognitive load classification was carried out. The quality of classification was measured with F1 and accuracy, which provided similar results. The following classifiers have been tested: decision tree with the Gini splitting criterion and unlimited depth, kNN with k = 5, SVM with a quadratic kernel, SVM with a cubic kernel, SVM with a linear kernel (C = 1, gamma = 1/16, other values of hyperparameters were tested in the course of preliminary experiments; however, no notable influence of classification performance was registered), logistic regression, random forest with 100 trees with the Gini splitting criterion as well as multilayer perceptron with two hidden layers and the ReLU activation function. The best obtained results are presented in Table 1. The first column contains the name of the classifier and the second contains the average accuracy score and the standard deviation in parentheses. The third column includes the mean values of the F1 measure and the standard deviation in parentheses. The fourth column shows the number of features used for the classification selected on the basis of the feature ranking. The column contains the number of features selected from the ranking. The number of features was the same in case of accuracy and the F1 measure. As shown, the best results were achieved for the random forest classifier—almost 96% with 16 features. Four classifiers obtained results above 90%. The SVM classifier with linear kernel needed first 10 features which led to the set of 14 features contained in the union of all separate classifier feature sets. The approach allowed us to obtain a score above 91%.

Table 1. Classification results for a selected feature subset.

Classifier	Accuracy	F1	Number of Features
Decision Tree	90.91 (6.73)	90.93 (6.72)	14
SVM With Linear Kernel	91.44 (6.62)	91.36 (6.72)	14
Logistic Regression	90.06 (7.58)	90.03 (7.68)	13
Random Forest	95.97 (4.55)	95.98 (4.52)	16

Figures 3–5 show how the accuracy of the classifier changes for the classifiers with respect to the number of features taken for analysis. Feature importance was obtained with logistic regression. The following models were tested: random forest, SVM with a linear kernel and decision tree. The standard deviation was marked as horizontal black lines. An interesting situation can be observed for the SVM classifier with a linear kernel. On the basis of just one feature, an accuracy of more than 82% can be obtained. As for the random forest classifier, the deviation value decreases with the increase of the feature numbers.

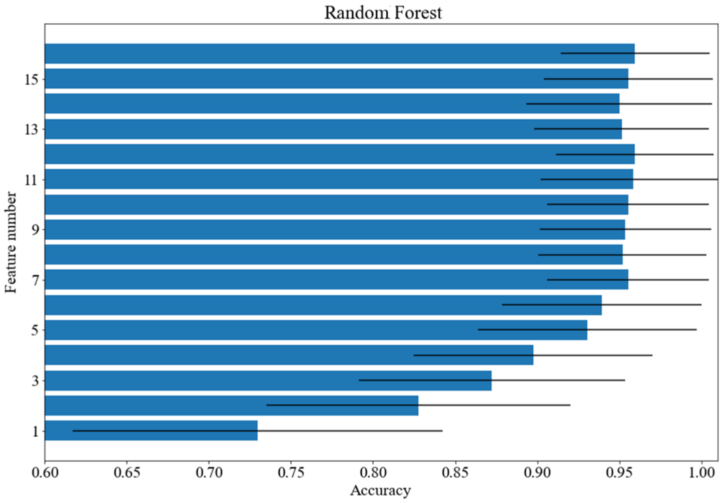


Figure 3. Accuracy scores for various number of features: random forest.

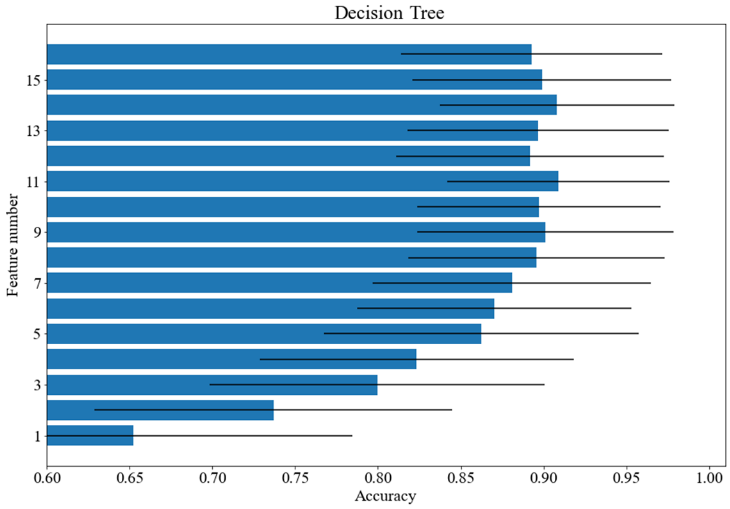


Figure 4. Accuracy scores for various number of features: decision tree.

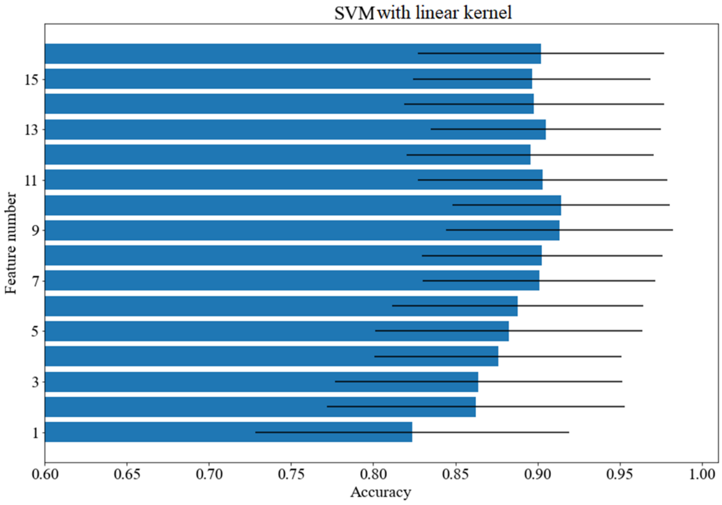


Figure 5. Accuracy scores for various number of features: SVM with linear kernel.

3.3. Feature Ranking

In addition to a quantitative influence on classification accuracy, interpretable machine learning methods allow us to obtain weights corresponding to the importance of particular features. The differences between feature importance weights can be quite low, hence a ranking containing sorted features can be an inadequate approach. A better approach would be to apply cluster analysis to divide features into three groups based on their importance: high, medium and low. The method of k-means was utilized for obtaining the three clusters. The presented values of weights were produced in the course of multiple repeated experiments including training a classifier model on a random sample drawn from the analysed dataset. The experiment was repeated 1000 times for each classifier model.

Figure 6 presents barplots of sorted feature importance weights divided into the mentioned groups. Each subplot corresponds to a separate level of cognitive workload. The k-means clustering was applied independently to separate workload levels. Two models were applied for calculating the features' weights: linear SVM and logistic regression. The models have been chosen as the main feature selectors due to the fact that they are among the most popular interpretable machine learning algorithms [45,46]. A ranking has been created for each of the levels: low, medium and high. The higher the feature is in the ranking, the more important it is. It can be noticed that some features from the low and medium levels are common, just like those from the medium and high levels.

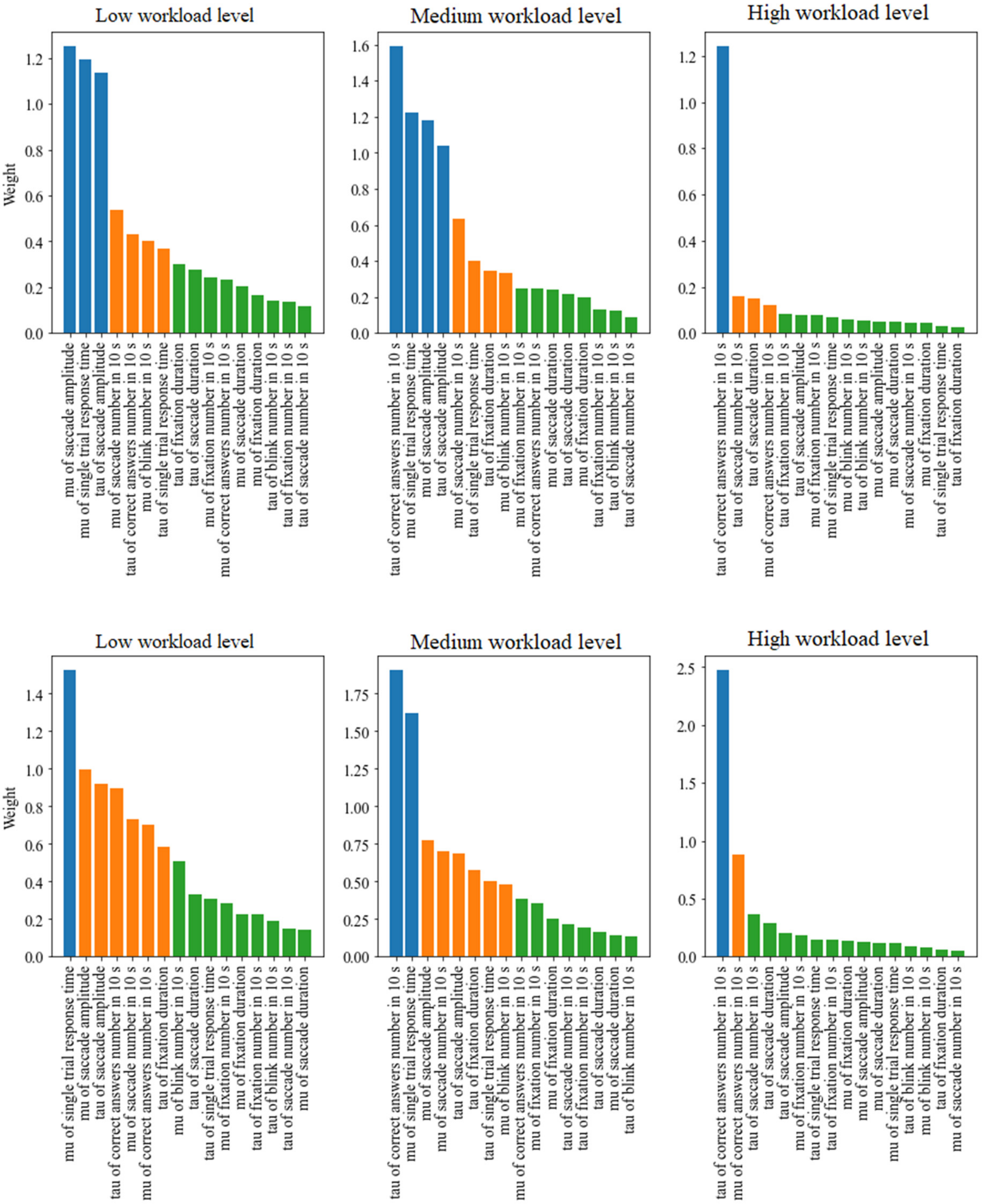


Figure 6. Clustered feature importance sorted in descending order. Top: SVM with linear kernel, bottom: logistic regression. Blue: high importance, orange: medium importance, green: low importance.

Table 2 presents the features belonging to the clusters of high and medium importance for all the levels of cognitive workload with the application of the linear SVM model. As Figure 6 shows, the subsets belonging to the specific clusters obtained by different models are very similar; the only differences are related to the following features: tau of fixation duration and mu of correct answers number in 10 s, which were included into the clusters of lower importance. Table 2 presents the features included in the clusters of high and medium importance by the SVM model with linear kernel. The detailed list of features belonging to the medium and high importance clusters are presented separately for linear SVM and logistic regression in supplementary materials in Tables S1 and S2.

Table 2. The features belonging to the clusters of high and medium importance according to linear SVM model, presented separately for each cognitive workload level.

Low Cognitive Workload	Medium Cognitive Workload	High Cognitive Workload
mu of blink number in 10 s mu of saccade amplitude mu of saccade number in 10 s mu of single trial response time tau of correct answers number in 10 s tau of saccade amplitude tau of single trial response time	mu of blink number in 10 s mu of saccade amplitude mu of saccade number in 10 s mu of single trial response time tau of correct answers number in 10 s tau of fixation duration tau of saccade amplitude tau of single trial response times	mu of correct answers number in 10 s tau of correct answers number in 10 s tau of saccade duration tau of saccade number in 10 s

When analysing the importance of the features in Table 2, it can be noticed that the common feature for all workload levels is tau of correct number of answers in 10 s. As for the feature sets selected for low and medium cognitive workload, they overlap to a high degree. The most important eye-tracking-related features are mu of blink number in 10 s, mu of saccade amplitude, mu of saccade number in 10 s and tau of saccade amplitude, whereas the set of the most important cognitive measures includes mu of single trial response time, tau of correct answers number in 10 s and tau of single trial response times.

Furthermore, a high cognitive workload has the set of distinct features that overlaps with the other levels to a notably lesser extent, which can be related to different cognitive processes corresponding to that level.

4. Discussion

The major aims of the study were to find out whether the cognitive and physiological data collected during cognitive-workload-related experiments fit the ex-Gaussian distribution, and to verify whether it is possible to efficiently perform three-class classification of cognitive workload levels using interpretable machine learning models. An independent approach was applied in the study, so that data from particular subjects were not separated into trial and the main test datasets. Such an approach ensures a better reflection of reality as in practice the trained model is usually applied to data that are taken entirely (without data division) from a new participant. An additional purpose was to obtain the feature interpretability, which is especially important in subject-independent classification.

The dataset applied in the classification was composed of two sets of features: eye-tracking and cognitive-based. Collected eye-tracking data related to such measurements as saccades, fixations and blinks; in other words, typical outputs gathered during an eye-tracking procedures. Additionally, cognitive features were gathered as a consequence of the DSST performance. Participants did not have any additional equipment; no devices had been directly attached to their bodies and did not restrict their movements or hinder their testing.

The feature extraction procedure was performed in such a way that features such as number of saccades, fixations or blinks and number of correct responses were extracted for 10 s intervals. The feature extraction procedure was focused on ex-Gaussian statistics, especially the mu and tau parameters. The sigma parameters were included in the initial analyses because most sigma parameters calculated for particular measures occurred to be insignificant; therefore, we discarded all the sigma parameters from further analysis.

An interpretable machine learning model was adapted in order to calculate the ranking of the most valuable features. This ranking was used to improve the classification results. Furthermore, it gives valuable information about the process of mental workload. Two models were applied to assign and to verify specific weights to separate features. These models are logistic regression with elastic net regularization and SVM with linear kernel.

Two models were chosen in order to verify the stability of feature importance ranking. It allows the placing of feature importance quantitatively in the model.

The prominence of features can be examined in two ways. The first is an approach including the whole set of features, cognitive and eye-tracking measures together. The feature ranking was analysed separately for each cognitive workload level (low, medium and high), as presented in Figure 6. K-means cluster analysis was applied to divide features into three groups of different importance: high, medium and low. Further analysis of feature importance was conducted based on features from the first two clusters, rejecting the third due its features having the lowest importance. Specific weights of features are presented in Figure 6. In our opinion considering features of the third cluster might be confusing. The detailed list of features belonging to particular clusters separately for linear SVM and logistic regression are presented in supplementary materials in Tables S1 and S2. The most important feature for all three levels turned out to be the tau of the correct number of answers (in 10 s). The most noticeable differences between the low and medium levels were the mu of single trial response time, the mu of saccade amplitude, the tau of saccade amplitude and the mu of saccade number (in 10s). The difference between the medium and high levels is particularly strong for the feature tau of the correct number of answers (in 10 s). Moreover, the tau of saccade amplitude is also placed high in the ranking. The second approach assumed the division into cognitive and eye-tracking features. The detailed ranking for these two sets is presented separately in supplementary materials (Tables S3 and S4 for logistic regression and Tables S5 and S6 for linear SVM). The cognitive feature set contained four features; the most important of them was the tau of the correct number of answers (in 10 s). In the case of the eye-tracking feature set including 12 features, the mu of saccade amplitude and the tau of saccade amplitude were high in the ranking. Overall results of feature ranking indicate features related to saccades, the number of correct answers and single trial response time as the most valuable in the case study. Results show that features related to fixations and blinks were not as important as other eye-tracking features, especially saccades.

The results of the classification show that the features based on ex-Gaussian statistics allow us to carry out a multi-class classification. Eight classifiers were initially tested and four of them enabled us to achieve an accuracy higher than 90%. They are logistic regression, random forest, SVM with a linear kernel and decision tree. Obtaining more than 90% accuracy is possible with 13 features (logistic regression). However, obtaining a result equal to almost 96% is possible with the use of 16 features. This result was obtained for the random forest classifier.

In sum, we find that introducing ex-Gaussian distributional characteristics to cognitive and physiological data associated with different levels of cognitive workload classification was beneficial. This claim seems to be particularly justified considering that among eye-tracking-related variables with the highest classification powers the majority covers the tau parameter, for example, considering saccade number, saccade duration and its amplitude. This means that a distributional feature which is usually deleted when analysed according to a parametric approach turns out to distinguish to the greatest extent three levels of cognitive processing overburden. However, the fact that mentioned indicators of eye-movement were characterised by the tau metric, and not by mu, suggests that cognitive workload, as it is studied with eye-tracker methodology, is probably highly associated with the amount of non-typical rare dimensions of eye-movement, not by the most common ones. Therefore, we suggest that the accumulation of cognitive workload might be physiologically expressed by an increase in the scope of outliers, not only by changes related to the most common features of performance.

There are some limitations of our study, which should be addressed. The experimental group was not well balanced regarding sex, with a clear predominance of men. However, according to our knowledge, there is no strong and unequivocal support for the notion that there are substantial sex differences regarding eye-movements schemata observed during non-sexual visual content scanning [47]. Nevertheless, we plan on providing a better gender-balanced group in future studies. Additionally, the feature ranking and classification rate might be changed in the case of a differently constructed experiment. It is worth examining the influence of the type and the order of tasks on the overall classification results. Changes in the requirements of the task which was administered to elicit a gradual increase in cognitive workload consisted in both elongation of the performance time (difference between parts 1, 2 and 3) and an increase in the number of stimuli to be processed (difference between part 1, 2 and 3). Therefore, our study did not distinguish whether the increase in cognitive load was generated by performance length or the scope of task internal complexity indicated by the number of given stimuli. However, our goal was not directly related to the problem of the relationship between the specificity of the task and the level of cognitive load. We assumed that the increasing modification of both the complexity of the task and its length should provoke an increase in cognitive overburden. The data achieved indicate the accuracy of our premise.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/brainsci12050542/s1, Table S1: The features belonging to the clusters of high and medium importance according to SVM with linear kernel, Table S2: The features belonging to the clusters of high and medium importance according to logistic regression, Table S3: Separate class feature rankings obtained by interpreting the weights of the logistic regression model with elastic net regularisation for cognitive features, Table S4: Separate class feature rankings obtained by interpreting the weights of the logistic regression model with elastic net regularisation for eye-tracking features, Table S5: Separate class feature rankings obtained by interpreting the weights of the linear SVM model for cognitive features, Table S6: Separate class feature rankings obtained by interpreting the weights of the linear SVM model for eye-tracking features, Table S7: Ex-Gaussian parameters describing the data used for classification.

Author Contributions: Conceptualisation, M.P.-W. and P.K.; methodology, M.P.-W. and P.K.; software, M.K. and M.T.; validation, M.K., M.T. and M.P.-W.; formal analysis, M.K., M.P.-W. and M.T.; data curation, M.K.; writing—original draft preparation, M.K., M.P.-W. and P.K.; writing—review and editing, M.K., M.P.-W., P.K. and M.T.; visualisation, M.K. and M.P.-W.; supervision, M.P.-W. and P.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted according to the guidelines of the ↔Declaration of Helsinki, and approved by the Ethics Committee of Lublin University of Technology↔(Approval ID 2/2018 from 19 October 2018).

Informed Consent Statement: Any informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chen, J.; Wang, H.; Wang, Q.; Hua, C. Exploring the fatigue affecting electroencephalography based functional brain networks during real driving in young males. *Neuropsychologia* **2019**, *129*, 200–211. [CrossRef]
2. Lobo, J.L.; Ser, J.D.; De Simone, F.; Presta, R.; Collina, S.; Moravek, Z. Cognitive workload classification using eye-tracking and EEG data. In Proceedings of the International Conference on Human-Computer Interaction in Aerospace, Paris, France, 14–16 September 2016; pp. 1–8.
3. Khushaba, R.N.; Kodagoda, S.; Lal, S.; Dissanayake, G. Driver drowsiness classification using fuzzy wavelet-packet-based feature-extraction algorithm. *IEEE Trans. Biomed. Eng.* **2010**, *58*, 121–131. [CrossRef]
4. Walter, C.; Wolter, P.; Rosenstiel, W.; Bogdan, M.; Spüler, M. Towards cross-subject workload prediction. In Proceedings of the 6th International Brain-Computer Interface Conference, Graz, Austria, 16–21 September 2014.
5. Thodoroff, P.; Pineau, J.; Lim, A. Learning robust features using deep learning for automatic seizure detection. In Proceedings of the Machine Learning for Healthcare Conference, Los Angeles, CA, USA, 19–20 August 2016; pp. 178–190.
6. Appel, T.; Scharinger, C.; Gerjets, P.; Kasneci, E. Cross-subject workload classification using pupil-related measures. In Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications, Warsaw, Poland, 14–17 June 2018; pp. 1–8.

7. Hefron, R.; Borghetti, B.; Schubert Kabban, C.; Christensen, J.; Estepp, J. Cross-participant EEG-based assessment of cognitive workload using multi-path convolutional recurrent neural networks. *Sensors* **2018**, *18*, 1339. [CrossRef]
8. Fridman, L.; Reimer, B.; Mehler, B.; Freeman, W.T. Cognitive load estimation in the wild. In Proceedings of the 2018 Chi Conferenceon Human Factors in Computing Systems, Montreal, QC, Canada, 21–26 April 2018; pp. 1–9.
9. Hajinoroozi, M.; Mao, Z.; Jung, T.P.; Lin, C.T.; Huang, Y. EEG-based prediction of driver's cognitive performance by deep convolutional neural network. *Signal Proc. Imag. Commun.* **2016**, *47*, 549–555. [CrossRef]
10. Wang, Z.; Hope, R.M.; Wang, Z.; Ji, Z.; Gray, W.D. Cross-subject workload classification with a hierarchical Bayes model. *NeuroImage* **2012**, *59*, 64–69. [CrossRef]
11. Jimnez-Guarneros, M.; Gomez-Gil, P. Custom Domain Adaptation: A new method for cross-subject, EEG-based cognitive load recognition. *IEEE Sign. Proc. Let.* **2020**, *27*, 750–754. [CrossRef]
12. Nuamah, J.K.; Seong, Y. Support vector machine (SVM) classification of cognitive tasks based on electroencephalography (EEG) engagement index. *Br. Comput. Interf.* **2017**, *5*, 1–12. [CrossRef]
13. Atasoy, H.; Yildirim, E. Classification of Verbal and Quantitative Mental Tasks Using Phase Locking Values between EEG Signals. *Int. J. Signal Process. Image Process. Pattern Recognit.* **2016**, *9*, 383–390. [CrossRef]
14. Yamada, Y.; Kobayashi, M. Detecting mental fatigue from eye-tracking data gathered while watching video: Evaluation in younger and older adults. *Artif. Intell. Med.* **2018**, *91*, 39–48. [CrossRef]
15. Zarjam, P.; Epps, J.; Chen, F.; Lovell, N.H. Estimating cognitive workload using wavelet entropy-based features during an arithmetic task. *Comput. Biol. Med.* **2013**, *43*, 2186–2195. [CrossRef]
16. Işbilir, E.; Çakır, M.P.; Acartürk, C.; Tekerek, A. Şimcek Towards a Multimodal Model of Cognitive Workload Through Synchronous Optical Brain Imaging and Eye Tracking Measures. *Front. Hum. Neurosci.* **2019**, *13*, 375. [CrossRef] [PubMed]
17. Ziegler, M.D.; Kraft, A.; Krein, M.; Lo, L.-C.; Hatfield, B.; Casebeer, W.; Russell, B. Sensing and Assessing Cognitive Workload Across Multiple Tasks. In Proceedings of the International Conference on Augmented Cognition, Toronto, ON, Canada, 17–22 July 2016; Springer Nature: Cham, Switzerland, 2016; pp. 440–450.
18. Almogbel, M.A.; Dang, A.H.; Kameyama, W. EEG-signals based cognitive workload detection of vehicle driver using deep learning. In Proceedings of the 2018 20th International Conference on Advanced Communication Technology (ICACT), Chuncheon, Korea, 11–14 February 2018; pp. 256–259.
19. McKendrick, R.; Feest, B.; Harwood, A.; Falcone, B. Theories and Methods for Labeling Cognitive Workload: Classification and Transfer Learning. *Front. Hum. Neurosci.* **2019**, *13*, 295. [CrossRef]
20. Luce, R.D. *Response Times: Their Role in Inferring Elementary Mental Organization (No. 8)*; Oxford University Press on Demand: Oxford, UK, 1986.
21. Matzke, D.; Wagenmakers, E.J. Psychological interpretation of the ex-Gaussian and shifted Wald parameters: A diffusion model analysis. *Psychon. Bull. Rev.* **2009**, *16*, 798–817. [CrossRef] [PubMed]
22. Lacouture, Y.; Cousineau, D. How to use MATLAB to fit the ex-Gaussian and other probability functions to a distribution of response times. *Tutor. Quant. Methods Psychol.* **2008**, *4*, 35–45. [CrossRef]
23. Gmehlin, D.; Fuermaier, A.B.; Walther, S.; Debelak, R.; Rentrop, M.; Westermann, C.; Sharma, A.; Tucha, L.; Koerts, J.; Tucha, O.; et al. Intraindividual variability in inhibitory function in adults with ADHD—An ex-Gaussian approach. *PLoS ONE* **2014**, *9*, e112298. [CrossRef]
24. Thomson, D.R.; Seli, P.; Besner, D.; Smilek, D. On the link between mind wandering and task performance over time. *Conscious. Cogn.* **2014**, *27*, 14–26. [CrossRef]
25. Krukow, P.; Jonak, K.; Karpiński, R.; Karakuła-Juchnowicz, H. Abnormalities in hubs location and nodes centrality predict cognitive slowing and increased performance variability in first-episode schizophrenia patients. *Sci. Rep.* **2019**, *9*, 1–13. [CrossRef]
26. Pereira, E.J.; Gurguryan, L.; Ristic, J. Trait-Level Variability in Attention Modulates Mind Wandering and Academic Achievement. *Front. Psychol.* **2020**, *11*, 909. [CrossRef]
27. Robison, M.K.; Unsworth, N. Cognitive and contextual correlates of spontaneous and deliberate mind-wandering. *J. Exp. Psychol. Learn. Mem. Cogn.* **2018**, *44*, 85. [CrossRef]
28. Stawarczyk, D.; D'Argembeau, A. Conjoint influence of mind-wandering and sleepiness on task performance. *J. Exp. Psychol. Hum. Percept. Perform.* **2016**, *42*, 1587. [CrossRef]
29. Killingsworth, M.; Gilbert, T. A wandering mind is an unhappy mind. *Science* **2010**, *330*, 932. [CrossRef] [PubMed]
30. Moukhtarian, T.R.; Reinhard, I.; Morillas-Romero, A.; Ryckaert, C.; Mowlem, F.; Bozhilova, N.; Moran, P.; Ebner-Priemer, U.; Asherson, P. Wandering minds in attention-deficit/hyperactivity disorder and borderline personality disorder. *Eur. Neuropsychopharmacol.* **2020**, *38*, 98–109. [CrossRef] [PubMed]
31. Chen, H.; Nomi, J.S.; Uddin, L.Q.; Duan, X.; Chen, H. Intrinsic functional connectivity variance and state-specific underconnectivity in autism. *Hum. Brain Mapp.* **2017**, *38*, 5740–5755. [CrossRef] [PubMed]
32. Li, Y.; Zhu, Y.; Nguchu, B.A.; Wang, Y.; Wang, H.; Qiu, B.; Wang, X. Dynamic Functional Connectivity Reveals Abnormal Variability and Hyper-connected Pattern in Autism Spectrum Disorder. *Autism Res.* **2020**, *13*, 230–243. [CrossRef] [PubMed]
33. Koenig, J.; Kemp, A.H.; Feeling, N.R.; Thayer, J.F.; Kaess, M. Resting state vagal tone in borderline personality disorder: A meta-analysis. *Prog. Neuropsychopharmacol. Biol. Psychiatry* **2016**, *64*, 18–26. [CrossRef] [PubMed]
34. Spangler, D.P.; Williams, D.P.; Speller, L.F.; Brooks, J.R.; Thayer, J.F. Resting heart rate variability is associated with ex-Gaussian metrics of intra-individual reaction time variability. *Int. J. Psychophysiol.* **2018**, *125*, 10–16. [CrossRef]

35. Karakula-Juchnowicz, H.; Gałęcka, M.; Rog, J.; Bartnicka, A.; Łukaszewicz, Z.; Krukow, P.; Morylowska-Topolska, J.; Skonieczna-Zydecka, K.; Krajka, T.; Jonak, K.; et al. The food-specific serum IgG reactivity in major depressive disorder patients, irritable bowel syndrome patients and healthy controls. *Nutrients* **2018**, *10*, 548. [CrossRef]
36. Otero-Millan, J.; Troncoso, X.G.; Macknik, S.L.; Serrano-Pedraza, I.; Martinez-Conde, S. Saccades and microsaccades during visual fixation, exploration, and search: Foundations for a common saccadic generator. *J. Vis.* **2008**, *8*, 21. [CrossRef]
37. Guy, N.; Lancry-Dayan, O.C.; Pertzov, Y. Not all fixations are created equal: The benefits of using ex-Gaussian modeling of fixation durations. *J. Vis.* **2020**, *20*, 9. [CrossRef]
38. Boake, C. From the Binet-Simon to the Wechsler-Bellevue: Tracing the history of intelligence testing. *J. Clin. Exp. Neuropsychol.* **2002**, *24*, 383–405. [CrossRef]
39. Rayner, K. Eye movements in reading and information processing: 20 years of research. *Psychol. Bull.* **1998**, *124*, 372–422. [CrossRef] [PubMed]
40. Hessels, R.S.; Niehorster, D.C.; Nyström, M.; Andersson, R.; Hooge, I.T. Is the eye-movement field confused about fixations and saccades? A survey among 124 researchers. *R. Soc. Open Sci.* **2018**, *5*, 180502. [CrossRef] [PubMed]
41. Tobii, A.B. Tobii Studio User's Manual. Available online: https://www.tobiipro.com/siteassets/tobii-pro/user-manuals/tobii-pro-studio-user-manual.pdf (accessed on 7 October 2020).
42. Krukow, P.; Szaniawska, O.; Harciarek, M.; Plechawska-Wójcik, M.; Jonak, K. Cognitive inconsistency in bipolar patients is determined by increased intra-individual variability in initial phase of task performance. *J. Affect. Disord.* **2017**, *210*, 222–225. [CrossRef] [PubMed]
43. Krukow, P.; Harciarek, M.; Morylowska-Topolska, J.; Karakuła-Juchnowicz, H.; Jonak, K. Ineffective initiation contributes to deficient verbal and non-verbal fluency in patients with schizophrenia. *Cogn. Neuropsychiatry* **2017**, *22*, 391–406. [CrossRef] [PubMed]
44. Krukow, P.; Harciarek, M.; Grochowski, C.; Makarewicz, A.; Jonak, K.; Karakuła-Juchnowicz, H. What specifically contributes to disturbed non-verbal fluency in patients with bipolar disorder: Ineffective performance initiation, slowed processing or lack of the execution strategy? *Psychiatry Res.* **2019**, *271*, 15–22. [CrossRef]
45. Du, M.; Liu, N.; Hu, X. Techniques for interpretable machine learning. *Commun. ACM* **2019**, *63*, 68–77. [CrossRef]
46. Khaire, U.M.; Dhanalakshmi, R. Stability of feature selection algorithm: A review. *J. King Saud Univ.-Comput. Inf. Sci.* **2019**, *34*, 1060–1073. [CrossRef]
47. Toth, A.J.; Campbell, M.J. Investigating sex differences, cognitive effort, strategy, and performance on a computerised version of the mental rotations test via eye tracking. *Sci. Rep.* **2019**, *9*, 19430. [CrossRef]

7 Scientific Profile of the Candidate

7.1 Education

2018 - currently **Polish Japanese Academy of Information Technology**

Field: Computer Science, PhD degree (in progress).
The average of the marks equals 5.

2013 - 2018 **Lublin University of Technology**

Field: Computer Science, Bachelor and Master degree.
Bachelor thesis: "*Interactive course of numerical analysis*" with excellent grade
Master thesis: "*Identification, characterization, and correction of artifacts in electroencephalographic data*" with excellent grade.

I was the student with the best grades at the year. The University offered to me the position of research and teaching assistant.

7.2 Professional Career

10.2022 - currently **Reckitt Benckiser**

Data Scientist in IT Hub: machine learning models implementing, data analyzing, natural language processing, and writing scientific papers.

06.2016 - 09.2022 **Lublin University of Technology**

- **Research and teaching assistant in Computer Science Department: 10.2018 – 09.2022**

  Scientific activities included various stages of oculography and EEG signal processing: data collection, preparation, data cleaning, feature extraction, preparing and testing machine learning models, classification and analysis of cognitive workload level (interpretable machine learning models, fuzzy aggregation of results obtained from model ensemble), brain computer interfaces, algorithms of extraction of characters from handwritten texts, statistical analysis of results in Python.

  Conducting academic courses: Numerical Analysis, Algorithms and Data Structures, Advanced Object-oriented programming in C++, SQL and NoSQL databases management, Universal design, Software Engineering

- **Internship: 02.2017 – 06.2018**

  Scientific activities were related to analysis of EEG signals, cognitive workload and brain computer interfaces using machine learning and statistical tools in Python.

7.3 Academic Achievements

Google Scholar	The link to Google Scholar: https://scholar.google.com/citations?user=4yNfNoIAAAAJ&hl=en My publications were cited 172 times and h-index equals 6.
Research Gate	The link to Research Gate: https://www.researchgate.net/profile/Monika-Kaczorowska My publications were citied 109 times, and h-index equals 5.
ORCID	The link to ORCID: https://orcid.org/0000-0002-4618-7937

I have participated in several academic projects for example:

1. The scientific project "Development and analysis of neural network based algorithms for handwriting recognition". Activities carried out in the project: designing and developing an application allowing to create a character database in Python, creating a database of handwritten characters, developing algorithms for detection of handwritten characters, verification of the created handwritten characters database.

2. "Mobile Application Development – Joint Master Studies", International cooperation with National University of Uzbekistan in Tashkent, Republic of Uzbekistan. Elaboration of a task book for the subject entitled "Scientific Methods in Research Experiments" in English. The materials include topics related to conducting statistical analysis with application of the R programming language.

3. Elaboration of teaching academic books. The books were dedicated to the following subjects: "Universal design", "Fundamentals of algorithms and data structures", "Advanced object-oriented programming in C++", "SQL and NoSQL databases management", "Appliances of databases", ”NoSQL databases" in Polish language.

I have taken part in international conferences presenting scientific papers for example:

1. The 2020 International Conference on Computational Intelligence, Information Technology and Systems Research, December 17-20, 2020, title of presentation "Explainable classification of low and high mental fatigue with eye-tracking features".

2. 14th annual International Technology, Education and Development Conference, INTED 2020, 2-4 March 2020, Valencia, title of presentation: "Is it a good idea to learn numerical methods using mobile phone?" and "Gamification as a tool promoting a pro-ecological attitude among drivers?".

3. Modern Computational Methods and their Applications in Engineering Science, 24-25.11.2020, Lublin, title of presentation: "User experience driven assessment of selected features and perspectives of instant messaging applications"

7.3.1 List of all publications

I am the author of 30 publications, which are listed below. Each scientific paper has assigned points according to the Ministry of Science and Higher Education (MSHE).

1. Plechawska-Wójcik, M., Augustynowicz, P., **Kaczorowska, M.**, Zabielska-Mendyk, E., & Zapała, D. (2023). The Influence Assessment of Artifact Subspace Reconstruction on the EEG Signal Characteristics. *Applied Sciences, 13(3)*, 1605.

2. **Kaczorowska, M**., Plechawska-Wójcik, M., Tokovarov, M., & Krukow, P. (2022). Automated Classification of Cognitive Workload Levels Based on Psychophysiological and Behavioural Variables of Ex-Gaussian Distributional Features. *Brain Sciences, 12(5)*, 542, [MSHE: 100]

3. Dolecki, M., Karczmarek, P., Gałka, Ł., Plechawska-Wójcik, M., **Kaczorowska, M.**, Tokovarov, M., & Czerwinski, D. (2022, July). On the Understanding of Anomalies in the Oculography Data and Their Classification with an Application of Fuzzy Aggregators. *In 2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)* (pp. 1-6). IEEE, [MSHE: 140].

4. **Kaczorowska, M.**, Plechawska-Wójcik, M., & Tokovarov, M. (2021). Interpretable machine learning models for three-way classification of cognitive workload levels for eye-tracking features. *Brain sciences, 11(2)*, 210, [MSHE: 100].

5. **Kaczorowska, M.**, Karczmarek, P., Plechawska-Wójcik, M., & Tokovarov, M. (2021). On the Improvement of Eye Tracking-Based Cognitive Workload Estimation Using Aggregation Functions. *Sensors, 21(13),* 4542, [MSHE: 100].

6. Plechawska-Wójcik, M., Karczmarek, P., Krukow, P., **Kaczorowska, M.**, Tokovarov, M., & Jonak, K. (2021). Recognition of Electroencephalography-Related Features of Neuronal Network Organization in Patients With Schizophrenia Using the Generalized Choquet Integrals. *Frontiers in Neuroinformatics, 63,* [MSHE: 140].

7. Lukasik, E., Charytanowicz, M., Milosz, M., Tokovarov, M., **Kaczorowska, M.**, Czerwinski, D., & Zientarski, T. (2021). Recognition of handwritten Latin characters with diacritics using CNN. *Bulletin of the Polish Academy of Sciences. Technical Sciences, 69(1),* [MSHE: 100].

8. Tokovarov, M., **Kaczorowska, M**., & Miłosz, M. (2020). Development of Extensive Polish Handwritten Characters Database for Text Recognition Research. Advances in Science and Technology. *Research Journal, 14(3),* [MSHE: 100].

9. **Kaczorowska, M.**, Wawrzyk, M., & Plechawska-Wójcik, M. (2020, October). Binary Classification of Cognitive Workload Levels with Oculography Features. *In International Conference on Computer Information Systems and Industrial Management* (pp. 243-254). Springer, Cham., [MSHE: 40].

10. **Kaczorowska, M.** (2020). Analysis of typical programming mistakes made by first and second year IT students. *Journal of Computer Sciences Institute, 15,* [MSHE: 5].

11. Plechawska-Wojcik, M., Drozdzyk, T., **Kaczorowska, M.**, & Tokovarov, M. (2020). Gamification as A Tool Promoting a Pro-Ecological Attitude among Drivers. *In INTED2020 Proceedings* (pp. 6091-6097). IATED, [MSHE: 5].

12. **Kaczorowska, M.**, Tokovarov, M., Panczyk, B., & Plechawska-Wojcik, M. (2020). IS IT A GOOD IDEA TO LEARN NUMERICAL METHODS USING MOBILE PHONE?. *In INTED2020 Proceedings* (pp. 5869-5877). IATED, [MSHE: 5].

13. **Kaczorowska, M.**, & Tokovarov, M. USER EXPERIENCE DRIVEN ASSESSMENT OF SELECTED FEATURES AND PERSPECTIVES OF INSTANT MESSAGING APPLICATIONS. Monografie–Politechnika Lubelska, 62, [MSHE: 20].

14. Plechawska-Wójcik, M., Tokovarov, M., **Kaczorowska, M.**, & Zapała, D. (2019). A three-class classification of cognitive workload based on EEG spectral data. Applied *Sciences, 9(24),* 5340, [MSHE: 100].

15. Plechawska-Wójcik, M., **Kaczorowska, M.**, & Michalik, B. (2019). Comparative analysis of two-group supervised classification algorithms in the study of P300-based brain-computer interface. *In MATEC Web of Conferences* (Vol. 252, p. 03010). EDP Sciences, [MSHE: 5].

16. Tokovarov, M., Plechawska-Wójcik, M., & **Kaczorowska, M.** (2019, June). Multi-class classification of EEG spectral data for artifact detection. *In International Conference on Artificial Intelligence and Soft Computing* (pp. 305-316). Springer, Cham., [MSHE: 20].

17. Plechawska-Wojcik, M., **Kaczorowska, M.**, & Zapala, D. (2018, September). The artifact subspace reconstruction (ASR) for EEG signal correction. A comparative study. *In International Conference on Information Systems Architecture and Technology* (pp. 125-135). Springer, Cham., [MSHE: 20].

18. Plechawska-Wojcik, M., Borys, M., Tokovarov, M., **Kaczorowska, M.**, Wesolowska, K., & Wawrzyk, M. (2018, July). Classifying Cognitive Workload Based on Brain Waves Signal in the Arithmetic Tasks' Study. *In 2018 11th International Conference on Human System Interaction (HSI)* (pp. 277-283). IEEE, [MSHE: 15].

19. Plechawska-Wójcik, M., Borys, M., Tokovarov, M., & **Kaczorowska, M.** (2017, September). Measuring Cognitive Workload in Arithmetic Tasks Based on Response Time and EEG Features. *In International Conference on Information Systems Architecture and Technology* (pp. 59-72). Springer, Cham., [MSHE: 20].

20. **Kaczorowska, M.**, Pańczyk, B., & Dmytruk, R. (2018). Satisfaction of it students in numerical methods learning using educational application–research results. *In INTED2018 Proceedings* (pp. 4168-4175). IATED, [MSHE: 15].

21. Plechawska-Wójcik, M., Wesołowska, K., Wawrzyk, M., **Kaczorowska, M.**, & Tokovarov, M. (2017). Analysis of applied reference leads influence on an EEG spectrum. *Informatyka, Automatyka, Pomiary w Gospodarce i Ochronie Środowiska, 7(2),* 44-49, [MSHE: 7].

22. Plechawska-Wójcik, M., Wawrzyk, M., Wesołowska, K., **Kaczorowska, M.**, Tokovarov, M., Dmytruk, R., & Borys, M. (2017). EEG spectral analysis of human cognitive workload study. *Studia Informatica, 38(2),* 17-30, [MSHE: 9]

23. **Kaczorowska, M.** (2017). Identification, characterisation, and correction of artefacts in electroencephalographic data in study of stationary and mobile electroencephalograph. In ITM Web of Conferences (Vol. 15, p. 01003). *EDP Sciences,* [MSHE: 15]..

24. Borys, M., Tokovarov, M., Wawrzyk, M., Wesołowska, K., Plechawska-Wójcik, M., Dmytruk, R., & **Kaczorowska, M.** (2017, March). An analysis of eye-tracking and electroencephalography data for cognitive load measurement during arithmetic tasks. *In 2017 10th International Symposium on Advanced Topics in Electrical Engineering (ATEE)* (pp. 287-292). IEEE, [MSHE: 15].

25. **Kaczorowska, M.** (2017, December). Comparative analysis of measures of biological artefacts identification in the electroencefalography signal in the context of biological noise. *In 2017 International Conference on Electromagnetic Devices and Processes in Environment Protection with Seminar Applications of Superconductors (ELMECO & AoS)* (pp. 1-4). IEEE, [MSHE: 15].

26. **Kaczorowska, M.**, Plechawska-Wojcik, M., Tokovarov, M., & Dmytruk, R. (2017, March). Comparison of the ICA and PCA methods in correction of EEG signal artefacts. *In 2017 10th International Symposium on Advanced Topics in Electrical Engineering (ATEE)* (pp. 262-267). IEEE, [MSHE: 15].

27. **Kaczorowska, M.**, Dmytruk, R., & Pańczyk, B. (2017). Interactive application supporting numerical methods teaching. *In INTED2017 Proceedings* (pp.536-545). IATED, [MSHE: 15].

28. Tokovarov, M., **Kaczorowska, M.**, & Plechawska-Wójcik, M. (2017, December). Towards human identification based on SSVEP response: A proof of concept study. *In 2017 International Conference on Electromagnetic Devices and Processes in Environment Protection with Seminar Applications of Superconductors (ELMECO & AoS)* (pp. 1-4). IEEE, [MSHE: 15].

29. Plechawska-Wójcik, M., & **Kaczorowska, M.** (2016). Performance analysis and optimal parameter selection for 300-based brain-computer interface. *Studia Informatica, 37(1),* 41-54., [MSHE: 9].

30. Plechawska-Wojcik, M., & **Kaczorowska, M.** (2016). TOWARDS STUDENT CONCENTRATION ASSESSMENT USING P300-BASED BRAIN-COMPUTER INTERFACE. *In INTED2016 Proceedings* (pp. 844-852). IATED, [MSHE: 15].

7.3.2 Honors and awards

1. Scientific Scholarship of the Minister of Science and Higher Education in the academic year 2016/2017, 2017/2018.

2. Award of the Rector of Lublin University of Technology in the academic year 2019/2020 for scientific activity.

3. Scientific Scholarship of the President of Lublin in the academic year 2020/2021.