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.

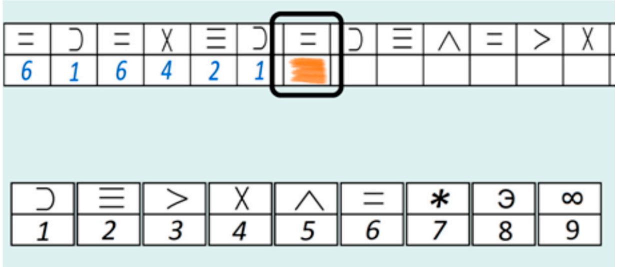

Figure 1. The interface of the application.

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

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

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

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

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

where $\{A_n\}$; n = 1, 2, . . . , 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, . . . , 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.

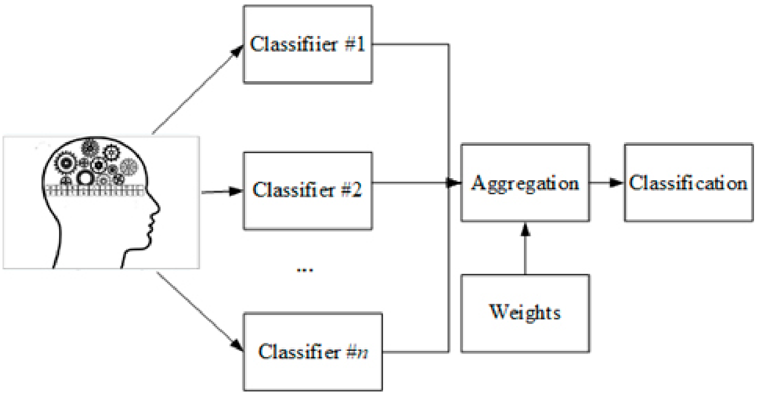

Figure 2. A general aggregation scheme.

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.

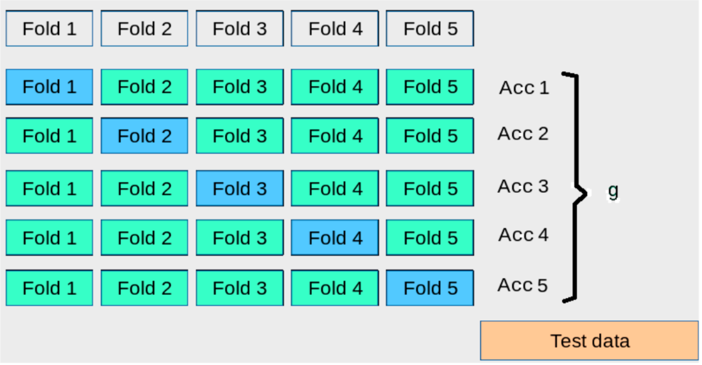

Figure 3. The idea of fuzzy measure generation through the process of cross validation.

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.

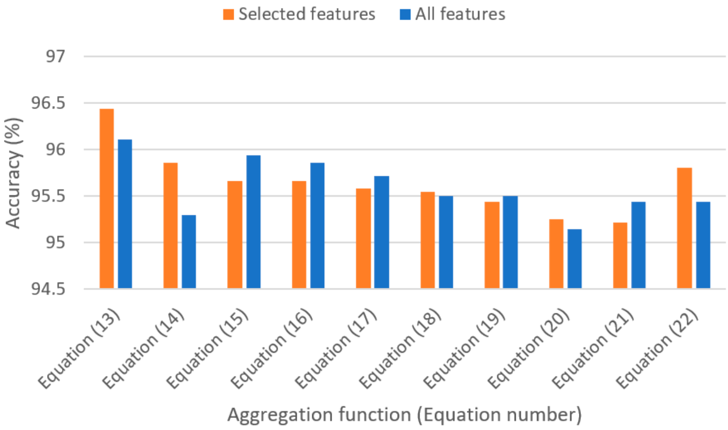

Figure 4. Comparison of accuracies achieved with top aggregation functions on complete feature set and for selected features.

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 ϵ [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]