# The Impact of Data Aggregation: Advocating for Individualized Analysis in Wearable Sensor Research

JEMMA L KÖNIG, JASCHA PENAREDONDO, NICK LIM, ANNIKA HINZE, and JUDY BOWEN, University of Waikato, New Zealand

The use of wearable sensors to record physiological data is becoming increasingly common, both in research and in daily life. User studies collect data using wrist bands, chest straps, and headbands to measure heart rate, skin conductance, and brain activity, to name a few. These readings can then be used to classify a range of physical and cognitive functions, such as cognitive and physical workload, cognitive and physical fatigue, stress, and attention. However, while physiological data is highly individualized, many researchers use machine learning classification methods that combine the participants' data into one dataset. In this paper, we demonstrate the negative impact this has on results and show that the individualized nature of physiological data requires individualized analysis and classification.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing**; **Visualization**; • **Computing methodologies → Machine learning**; • **Applied computing → Life and medical sciences**.

Additional Key Words and Phrases: Wearable technology, Participatory studies, Individualised data, Physiological data, Cognitive workload, Machine learning

**ACM Reference Format:**
Jemma L König, Jascha Penaredondo, Nick Lim, Annika Hinze, and Judy Bowen. 2018. The Impact of Data Aggregation: Advocating for Individualized Analysis in Wearable Sensor Research. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX).* ACM, New York, NY, USA, 16 pages. https://doi.org/XXXXXXX.XXXXXXX

## 1 Introduction

With recent advancements in wearable technology, both researchers and industry have begun investigating the use of physiological data to identify and predict cognitive functions and physical conditions. Measures such as brain activity and heart rate have been identified as suitable indicators of cognitive load (e.g., to identify driver workload [46]). The technology needed to collect appropriate data, as well as the AI and machine learning algorithms needed to make predictions from the data have become more accessible outside of specialist settings. This has enabled HCI researchers to use indicators of cognitive load as a measure of usability and accessibility of proposed systems, for example Novak et al.'s work on measuring cognitive load of virtual reality systems [47]. However, using physiological measures and cognitive load in user studies is not without it's challenges. Kosch et al. [36] noted that with the wide variety of methods and measures available, researchers may easily misuse methods or apply them out of context, leading to invalid results. Similarly, there is an increasing awareness of both ethics and effectiveness of such methods across a variety of

Authors' Contact Information: Jemma L König, jemma.konig@waikato.ac.nz; Jascha Penaredondo; Nick Lim; Annika Hinze; Judy Bowen, University of Waikato, Hamilton, New Zealand.

domains [35, 45]. Here we focus on the implications of combining data from multiple participants to identify patterns in individuals.

Our research focuses on the use of physiological readings to identify cognitive functions. However, physiological readings are highly personalised [12]. A resting heart rate of 40 beats per minute, for example, may be normal for one person, but of serious concern for another. For commercial wearable devices, like the Readiband Fatigue Predictor,[1] this is not a problem as all data collected belongs to one person and is collected over time, enabling a large personalised dataset to be collected and used for baseline and benchmarking. Yet, when using physiological data to identify cognitive functions, researchers tend to combine individual data together to form a larger, more generalised dataset (see Section 2.3), to train and test machine learning algorithms. However, this practice assumes a commonality in physiological data between participants. We suggest instead that the highly personalised nature of physiological data requires analysis on an individual level, and that doing so will provide better and more meaningful findings.

This paper uses cognitive workload as a case study to illustrate the issue. Cognitive workload refers to the mental effort and resources required to perform a particular task or cognitive activity. Prolonged periods of high cognitive workload can cause cognitive fatigue, which in turn can cause accidents or injuries [26]. As such, many researchers have begun investigating the classification of cognitive workload, often using physiological data points such as heart rate, heart rate variability, skin conductance, and brain activity [27, 43, 60].

This paper describes a study in which physiological data was recorded for 26 participants while they undertook first a resting task and then a cognitively intensive task. This dataset has then been used to predict cognitive workload using machine learning algorithms and a selection of evaluation regimes. machine learning algorithms can be trained and tested using a number of evaluation regimes. When working with physiological data, the most common involves combining all participants' data together. However, other, less common, approaches include training on data from $n-1$ participants and testing on the remaining participant data (leave-one-out), or treating each participant as their own dataset. In this paper, we apply these three evaluation regimes, and discuss the results in light of the repercussions of treating participatory data collectively versus individually.

We begin with a literature review that was used to determine the physiological data points to use for cognitive workload classification, and the machine learning algorithms and evaluation regimes that are most commonly used. After this, we outline our case study, including the study methodology, the machine learning classification, and the evaluation approaches. Finally we discuss our findings, highlight the difference in results, and provide visualisations of individualised physiological data.

## 2 Literature Review

A systematic literature review was conducted to determine (1) the physiological data points that are commonly used in cognitive workload research, (2) the classification methods commonly used in cognitive workload research, and (3) the evaluation approaches that are commonly used in cognitive workload research.

As shown in Figure 1, the first 100 articles (sorted by relevance) were selected from Google Scholar when using the search term *'extracting "physiological data" to predict "cognitive workload" using "machine learning" techniques'*. 39 of these articles were excluded based on the selection criteria listed below.

(1) The article must include one or more forms of physiological data
(2) The article must focus on cognitive workload (as opposed to other cognitive functions).

[1] https://fatiguescience.com/how-it-works/

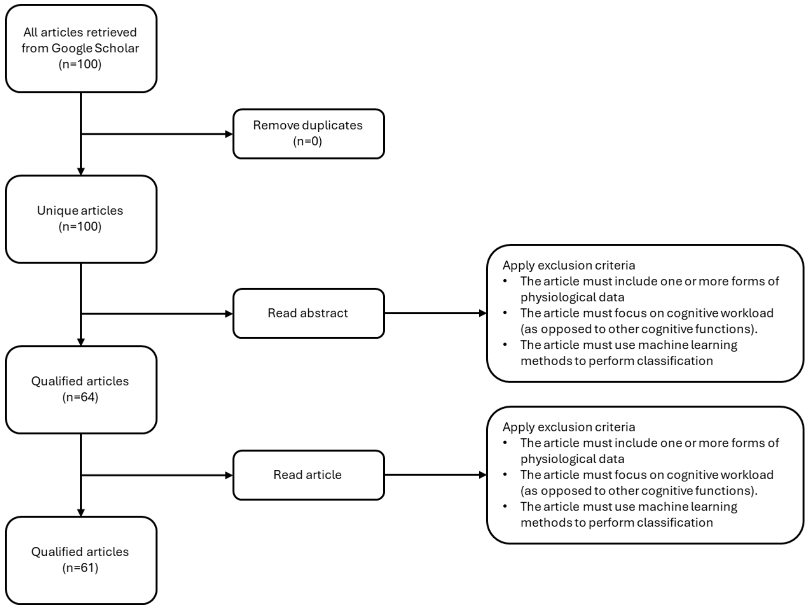



Fig. 1. Systematic literature review decision flow chart

Table 1. Physiological data points that are commonly used in cognitive workload research

| Physiological data points | Total number of articles | Citations |
|---|---|---|
| ECG | 38 | [3–5, 8–11, 13, 16–18, 20, 21, 24, 27, 28, 31, 32, 37, 38, 40–44, 46, 48, 50–53, 56, 58, 59, 64, 65, 70, 71] |
| EEG | 31 | [6, 7, 14, 16, 19, 20, 22–25, 28–31, 33, 39, 41, 46, 48, 53, 54, 57, 60, 62, 63, 65–68, 72, 73] |
| EDA | 26 | [4, 9–11, 13, 17–19, 24, 27, 28, 31, 37, 38, 40, 43, 44, 50, 52, 58, 59, 64, 65, 69–71] |

(3) The article must use machine learning methods to perform classification

In addition to the above selection criteria, the included articles had to be written in English, and include full text availability. This resulted in 61 articles being included in the final literature review.

### 2.1 Physiological data points

Table 1 illustrates the physiological data points we found to be commonly used in the literature. Electrocardiography (ECG) was the most commonly used physiological measure of cognitive workload, used in 38 articles. ECG records the heart's electrical activity, and is used to calculate Heart Rate (HR) and Heart Rate Variability (HRV). Electroencephalogram (EEG) was used in 31 articles. EEG measures the electrical activity of the brain, and was the second-most commonly used physiological measure of cognitive workload. EEG includes up to 64 channels, which are used to measure the activity in different areas of the brain (e.g. occipital, temperal, frontal, etc.). Electrodermal activity (EDA) was used in 26 articles. EDA measures variation of the electrical activity from the sweat glands, which is observed as

Table 2. Machine learning classifiers that are commonly used in cognitive workload research

| Machine Learning Classifiers | Total Number of Articles | Citations |
|---|---|---|
| SVM (including SVC) | 36 | [3, 5–7, 13, 16–18, 20, 21, 24, 25, 27, 29, 33, 37, 38, 41, 43, 44, 46, 48, 50–53, 55, 57, 59, 60, 62, 63, 65, 69–71] |
| Random Forest | 19 | [5–7, 13, 18, 29, 37–40, 43, 44, 50, 52, 53, 60, 62, 65, 68] |
| K-Nearest Neighbour | 18 | [3–7, 13, 18, 38, 39, 41, 43, 48, 52, 53, 57, 58, 62, 65] |

Table 3. Machine learning evaluation techniques that are commonly used in cognitive workload research

| Methods of Evaluation | Total Number of Articles | Citations |
|---|---|---|
| All-participants | 42 | [3, 5–9, 11, 14, 16–20, 24, 27, 29, 31, 33, 38–41, 43, 44, 46, 48, 50–52, 54–56, 60, 62–66, 68–71] |
| Leave-one-out | 11 | [23, 29, 31, 32, 37, 42, 48, 56, 71–73] |
| Individual | 8 | [4, 17, 25, 40, 46, 52, 58, 66] |

changes in the electrical conductance of the skin. The most-commonly used artifact of EDA is the skin conductance response (SCR). Other, less commonly used physiological data points included respiration rate and accelerometer data.

### 2.2 Cognitive workload classification methods

Table 2 illustrates the machine learning classifiers that were commonly used in the literature. Support Vector Machine (SVM) was used in 36 articles. SVM is a supervised learning method that can perform linear and non-linear classification and regression. Random forest (RF) was used in 19 articles. RF is an ensemble-learning technique in which many decision trees are used to provide solutions. K-Nearest Neighbour (KNN) was used in 18 articles. KNN is an instance-based supervised learning method. Upon classification of a new instance, KNN predicts based on the k-nearest training examples. Each of these are discussed further in Section 4.3. Finally, a collection of other machine learning classifiers were used in 15 or less articles each: Naive Bayes, Decision Tree, Logistic Regression, LDA, Neural Network based algorithms, and AdaBoost.

### 2.3 Methods of evaluation

When working with physiological data, we consider three different evaluation regimes, namely: all-participants, leave-one-out, and individual evaluation. Table 3 illustrates these commonly used evaluation methods in the literature. The all-participants method was used in 42 articles. This method involves aggregating all participants' data into one big dataset. This dataset is then used with either cross-validation, or a test and train split. The leave-one-out method was used in 11 articles. This method sets aside one participant's data to be the testing set, and uses the rest of the participants' data to train the model. Finally, the individual method was used in 8 articles. This method involves selecting one participant and using only that participant's data to build and train the machine learning model. Each of these are discussed further in Section 4.4.

## 3 Methodology

The original goal of this project was to develop an optimal machine learning model that can accurately classify the cognitive workload level of an individual, either resting or cognitive, based on various physiological readings. Based on the results of the literature review, we identified ECG, EDA and EEG as the most common physiological data points. We have selected ECG and EDA for use in our study. EEG has been excluded due to it's more invasive nature, and it's limitations in non-laboratory based environments. In addition to this, we have included accelerometer data. Accelerometer data is included in the sensors used for this study (see Section 3.2) and has been shown to be beneficial in the measurement of cognitive and physical workload [34].

### 3.1 Participants

The study was conducted with 26 participants between the ages of 20 and 40. Participants were university students who took part as a component of their undergraduate coursework. Ethical consent was received from the University ethics committee before the commencement of the study (HECS2023#06).

### 3.2 Equipment

The physiological measures recorded for each participant were HR, HRV, EDA, and accelerometer data (x, y, and z axes). Two wearable sensors were used throughout the study: (1) the Polar H10 Heart Rate Monitor, and (2) the Mindfield eSense Skin Response Sensor.

The Polar H10 Heart Rate Monitor[2] records raw ECG, HR, HRV, and accelerometer data. The sensor is attached using a chest strap, and was fitted around the torso of each participant. Data was collected via a mobile application called Reading People [35].

The Mindfield eSense Skin Response Sensor[3] records EDA. The sensor includes two electrodes, which were fitted on the middle and ring finger of the participant's non-dominant hand. This sensor was also interfaced with the Reading People application which transfers the aggregated EDA data via the headphone jack.

### 3.3 Protocol

The protocol for the study included two tasks: (1) a resting task, and (2) a cognitively-intensive task. Participants performed the study in a controlled environment (a quiet office space) to ensure a non-disruptive experience. Sessions were run at different times throughout the day and week. However, every effort was made to provide participants with a space devoid of external noise or distraction. The study was run as follows:

(1) The participant arrived and the study was explained to them. Participants were provided the opportunity to ask any questions and/or opt out of the study. Participants were provided with the ethics approval information, signed the ethical consent form and the study began.
(2) The sensors were fitted and the mobile application was set up to start recording.
(3) Task 1: Participants were asked to sit quietly and rest for 10 minutes. The start time was recorded.
(4) The cognitively intensive task was loaded onto the computer, and participants were shown how to use it.
(5) Task 2: Participants were asked to perform the cognitively intensive task for 10 minutes. The start time was recorded.

[2] https://www.polar.com/nz-en/sensors/h10-heart-rate-sensor
[3] https://mindfield-esense.com/esense-skin-response/

(6) At the end of the session, the physiological recording was stopped and the sensors were removed. The dataset was then saved to the researcher's computer.

The cognitively intensive task was conducted using the NASA Multiple Attribute Task Battery. The NASA Multiple Attribute Task Battery (MATB) is a flight simulator that requires participates to monitor and track multiple tasks at once. This includes a monitoring task, a tracking task, an auditory task, and a resource management task. For the purpose of our study, the cognitive task was set up using OpenMATB, an open-source version of MATB, in which tracking, monitoring, and resource managing are simulated simultaneously [15]. This system has been shown to induce cognitive workload [49]. OpenMATB was run on a Dell Latitude laptop using a Logitech Extreme 3D Pro joystick.

Each participant completed the study twice, on two separate occasions, with a minimum interval of 24 hours between sessions.

## 4 Classification

### 4.1 Data Aggregation and Extraction

As discussed in Section 3, a dataset of HR, HRV, EDA, and accelerometer data was collected from 26 participants. This resulted in 52 folders of data,[4] each containing four data files (HR.txt, HRV.txt, EDA.txt, ACC.txt). Each of the data files has been read in and converted to a dataframe[5] in Python. A sliding window of was applied to aggregate the data for each participant. Sliding windows are used to account for the different sample rates when recording physiological data. The sliding window had a width of 60 seconds and a slide of 5 seconds.

Once the data was aggregated, the next step was to extract the data that was recorded during the 10 minute resting task, and during the 10 minute cognitively intensive task. Additionally, the first and last minute of data have been removed from each task, in order to mitigate any noise that may have been introduced at the start and end of the tasks. This resulted in four eight-minute time-blocks for each participant (two resting time-blocks, and two cognitive workload time-blocks). Finally, classification labels were added to each of the data points: 0 for cognitive and 1 for resting.

### 4.2 Pre-processing features

Pre-processing features involves transforming and preparing data before it is fed into a machine learning model. This is necessary to enhance the quality of data and consequently improve the model's performance. Our data was pre-processed in two ways: imputation, and feature scaling.

Imputation involves replacing missing values within the variables of the dataset with statistical measures such as the mean, median, etc. Since the individual physiological readings were recorded at different frequencies, there were time differences when the data was aggregated. This results in missing values, hence the need for imputation, in this case using the median strategy. This means that the missing data is replaced with another value based on the median of the features with the missing values, a process that occurs within the training set.

Feature scaling involves normalising the data. This means to scale the features (i.e. variables) to a similar range to prevent certain features from dominating others upon model training. In particular, standardisation is used in order to scale the features to have a mean of 0 and a standard deviation of 1. This is particularly important for distance-based algorithms such as K-Nearest Neighbours or gradient descent-based algorithms.

[4]Two folders for each participant, based on the two study sessions
[5]A dataframe is a 2D array of rows and columns.

Table 4. Results: All-participants classification (the classifier with the highest result is highlighted)

| | SVC | RF | KNN |
|---|---|---|---|
| All participants | **0.739** | 0.738 | 0.697 |

### 4.3 Classifier selection

Based on our literature review findings, the following machine learning classifiers were used to perform the cognitive workload classification: Support Vector Classifier, Random Forest, and K-Nearest Neighbors.

Support Vector Classifier (SVC), or linear Support Vector Machine (SVM) is a supervised machine learning algorithm used for binary classification (i.e. where the goal is to predict one of two possible outcomes, usually represented as 0 or 1) [1]. This algorithm finds the hyperplane that best separates the two classes in a feature space. The purpose of this hyperplane is to maximise the distance between the two classes and consequently minimise classification error. SVC uses support vectors (i.e. data points that are the closest to the hyperplane) to determine the best position of the hyperplane.

Random Forest (RF) is an ensemble-based machine learning algorithm based on the principle of bagging (bootstrap aggregating) and randomness [2]. This algorithm builds a collection of trees by selecting random subsets of data and features from the dataset and then calculates the average of the trees' predictions to produce a more accurate result. The number of trees (or estimators) is a hyperparameter that can be tuned to get the optimal results, in this case estimators=10. In our findings, we fond that this relatively small number gave us the optimal results.

K-Nearest Neighbor (KNN) is an instance-based machine learning algorithm used for classification tasks. When classifying a new instance, KNN identifies the k-nearest data points from the training set and assigns the majority class among the neighbouring data points to the new data point [7]. The value of k determines the number of neighbours to consider. For example, if k=3 then the algorithm considers the three closest data points to the new instance. It is important to choose the appropriate k-value as this can significantly affect the accuracy of the model. For example, a small k-value would not be ideal with a big dataset as it would not be representative of the whole data. As a result, the number of neighbors that will be used in this case is 3.

### 4.4 Evaluation selection

The three evaluation regimes that were used to assess the effectiveness of the cognitive workload classification models are (1) all participants, (2) leave-one-out, and (3) individual. This allows us to evaluate both the classifiers themselves, and the different evaluation techniques.

As explained in Section 2.3, the all-participants method involves aggregating all the participants' data into a single dataset. This technique was used with stratified 5-fold cross-validation. Since there is a wide variety of participants, we predict that combining all of the participants' data together would bring about a lot of variation due to the uniqueness of individuals. Hence, we were interested in how well the machine learning models would perform with this technique.

The leave-one-out method sets aside one participant's data to be the testing set, and uses the rest of the participants' data to train the model. This utilises the leave-one-out cross-validation method, wherein each data point is used as a validation set at least once, while the rest of the data serves as the training set. As previously mentioned, the more participants' involved, the greater the data variability present, which can significantly influence the model's performance.

Table 5. Results: Leave-one-out classification (the classifier with the highest result is highlighted for each participant)

| | SVC | RF | KNN |
|---|---|---|---|
| P1 | 0.715 | **0.728** | 0.663 |
| P2 | **0.729** | 0.728 | 0.686 |
| P3 | **0.744** | 0.740 | 0.679 |
| P4 | **0.737** | 0.701 | 0.665 |
| P5 | 0.720 | **0.725** | 0.695 |
| P6 | **0.720** | 0.717 | 0.669 |
| P7 | 0.744 | **0.748** | 0.674 |
| P8 | 0.727 | **0.735** | 0.665 |
| P9 | 0.733 | **0.745** | 0.669 |
| P10 | **0.745** | 0.714 | 0.695 |
| P11 | **0.755** | 0.718 | 0.724 |
| P12 | **0.734** | 0.726 | 0.684 |
| P13 | 0.701 | **0.736** | 0.696 |
| P14 | **0.713** | 0.709 | 0.677 |
| P15 | **0.731** | 0.709 | 0.684 |
| P16 | **0.717** | 0.692 | 0.666 |
| P17 | 0.715 | **0.716** | 0.669 |
| P18 | 0.726 | **0.730** | 0.713 |
| P19 | 0.710 | **0.735** | 0.676 |
| P20 | 0.740 | **0.742** | 0.717 |
| P21 | **0.741** | 0.730 | 0.703 |
| P22 | 0.744 | **0.758** | 0.715 |
| P23 | **0.747** | 0.688 | 0.677 |
| P24 | **0.740** | 0.704 | 0.699 |
| P25 | **0.752** | 0.745 | 0.706 |
| P26 | 0.688 | **0.715** | 0.686 |

Finally, the individual method involves selecting one participant and using only this participant's data in the machine learning model. This has been used with the stratified 5-fold cross-validation and evaluated using the test accuracy metric. We predicte that using only one participant's data would have less data variability because physiological signals are unique to each individual. Hence, training and testing with the data of the same participant should provide a more accurate classification performance.

## 5 Results

Tables 4, 5, and 6 show the results for each evaluation technique. First, Table 4 shows the result for the "all participants" evaluation method. For this evaluation method, it can be seen that SVC is the best performing classifier, with an accuracy of 73.9%. RF has a similar accuracy (73.8%) while KNN performed the worst (69.7%). Next, Table 5 shows the results for the "leave-one-out" evaluation technique. As can be seen by the rows in the table, the leave-one-out evaluation was performed on each participant, where the model was trained on all other participants and then tested on the target participant. For this evaluation method, SVC performed best for 14 of the participants (54% of participants), while RF performed best for 12 participants (45%), and KNN did not perform best for any participants. Next, Table 6 shows the results for the "individual" evaluation technique. Similar to the leave-one-out method, as can be seen by the rows in the table, the individual evaluation was performed on each participant. For the individual method, each participant's data

Table 6. Results: Individual classification (the classifier with the highest result is highlighted for each participant)

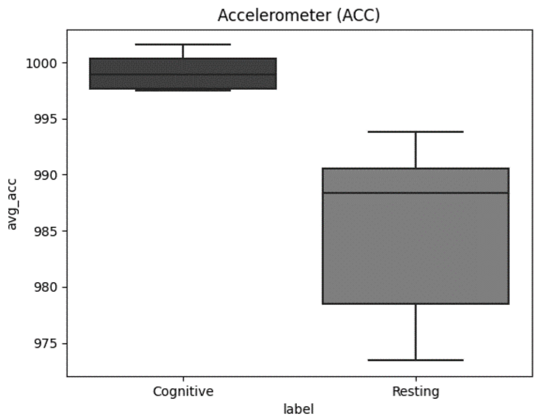
| | SVC | RF | KNN |
|---|---|---|---|
| P1 | **1.000** | **1.000** | **1.000** |
| P2 | 0.980 | 0.973 | **0.993** |
| P3 | 0.820 | **0.968** | 0.959 |
| P4 | 0.734 | **0.739** | 0.724 |
| P5 | **1.000** | **1.000** | 0.998 |
| P6 | **1.000** | **1.000** | **1.000** |
| P7 | **1.000** | **1.000** | **1.000** |
| P8 | **1.000** | **1.000** | **1.000** |
| P9 | **1.000** | **1.000** | **1.000** |
| P10 | **1.000** | **1.000** | **1.000** |
| P11 | **1.000** | **1.000** | 0.990 |
| P12 | **1.000** | **1.000** | 0.998 |
| P13 | **1.000** | **1.000** | **1.000** |
| P14 | **1.000** | **1.000** | **1.000** |
| P15 | **1.000** | **1.000** | **1.000** |
| P16 | **0.963** | 0.915 | 0.954 |
| P17 | **1.000** | 0.998 | **1.000** |
| P18 | **0.954** | 0.902 | 0.939 |
| P19 | **1.000** | **1.000** | **1.000** |
| P20 | **1.000** | 0.939 | 0.990 |
| P21 | 0.780 | **0.827** | 0.798 |
| P22 | 0.856 | **0.966** | 0.920 |
| P23 | 0.976 | 0.937 | **0.985** |
| P24 | **1.000** | 0.959 | 0.993 |
| P25 | 0.988 | 0.937 | **0.995** |
| P26 | 0.956 | **1.000** | 0.944 |

Table 7. Results: All classification methods (the classifier with the highest result is highlighted for each evaluation method)

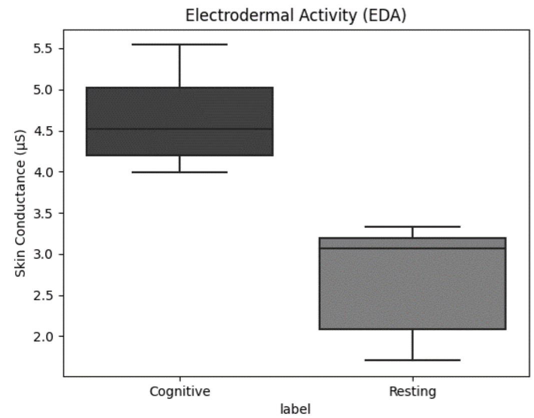
| | SVC | RF | KNN |
|---|---|---|---|
| All participants | **0.739** | 0.738 | 0.697 |
| Leave-one-out | **0.730** | 0.724 | 0.687 |
| Individual | 0.962 | 0.964 | **0.968** |

was treated individually and was trained and tested using cross-validation. For this evaluation method, both SVC and RF performed best (or best equal) for 18 of the participants, while KNN performed best (or best equal) for 14 participants.

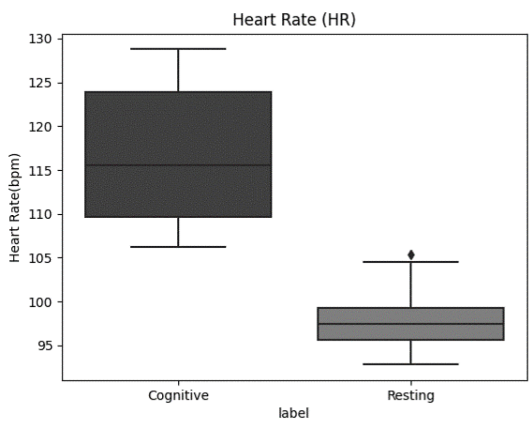
Finally, Table 7 shows the results averaged across participants. As can be seen here, and in Tables 4, 5, and 6, the individual evaluation method performed noticeably better than the all-participants and leave-one-out methods (~97% versus ~73%). This can be indicative of the highly individualised nature of physiological data (discussed further in Section 6). It should also be noted that KNN performs best when the individualised results are averaged across participants (shown in Table 7), while SVC and RF performed best when the results were considered on a per-participant basis (shown in Table 6).



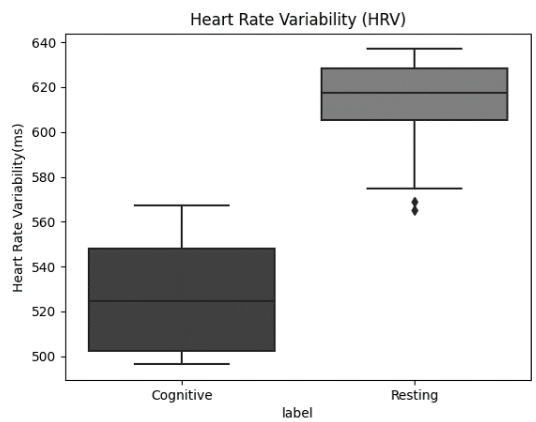

(a) ACC Readings




(b) EDA Readings




(c) HR Readings




(d) HRV Readings

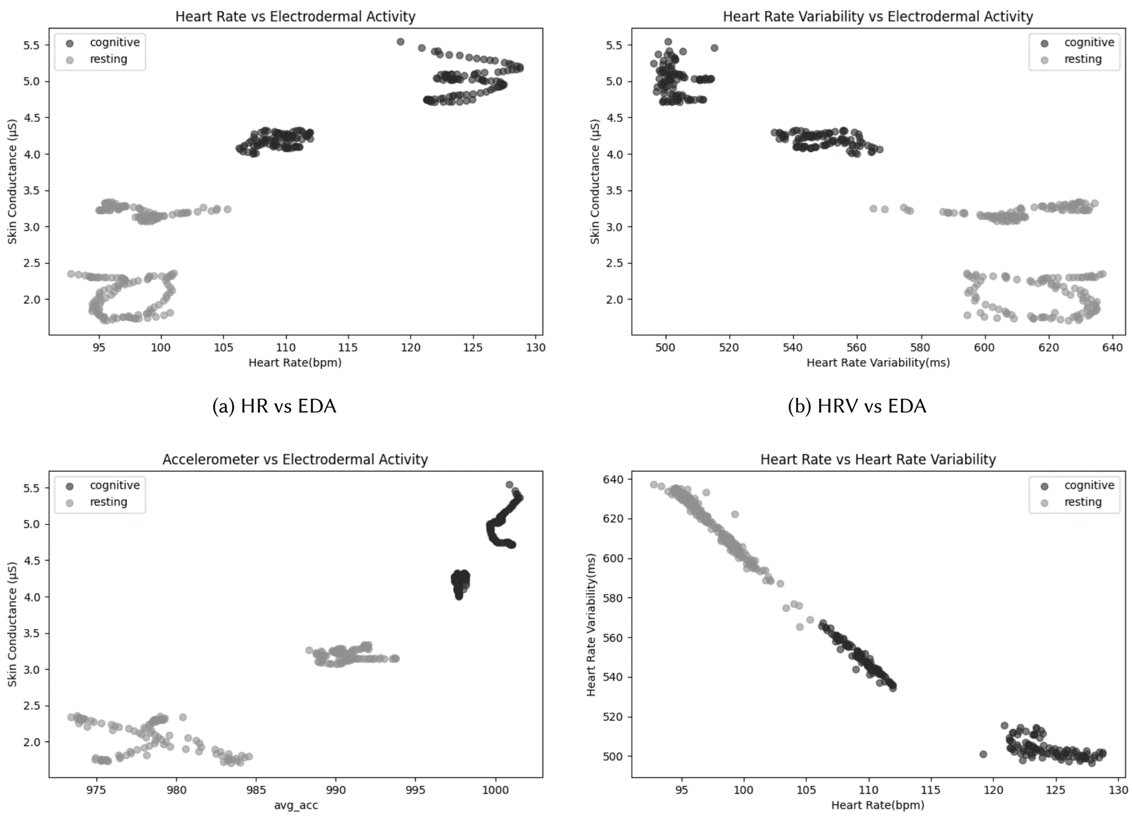
Fig. 2. Box plots of physiological readings (for one individual participant)

## 6 Discussion

The original goal of this project was to develop an optimal machine learning model that can accurately classify the cognitive workload level of an individual, either resting (class 1) or cognitive (class 0), based on various physiological readings. While we have successfully achieved this (with 97% accuracy), the most meaningful finding that this research produced was the difference in accuracy between evaluation regimes. Physiological data is highly personalised, and this was reflected in our results. All three models (SVC, RF, and KNN) produced higher accuracy (97%) when considering participants individually, as opposed to considering them collectively (69%-74%).

### 6.1 Participant visualisation

To further understand this pattern, we have visualised one of the participants data. Figure 2 shows a box plot for each of the data types: accelerometer (ACC), Electrodermal Activity (EDA), Heart Rate (HR), and Heart Rate Variability (HRV).




(a) HR vs EDA

(b) HRV vs EDA

(c) EDA vs ACC

(d) HR vs HRV

Fig. 3. Scatter plots of physiological readings (for one individual participant)

As seen in the figure, there is a clear separation between the cognitive and resting values of all four data types, notably so with the ACC and EDA readings. In terms of data variability, the ACC values have a very narrow spread for the cognitive class while the resting class has greater variability. A similar observation can be seen in the HR readings. Whereas, EDA and HRV have comparable variability for both the resting and cognitive classes.

Figure 3a also shows a clear separation between the data points of the resting class and the cognitive class. For instance, the resting class generally has low EDA and HR values while the cognitive class has higher EDA and HR values. Overall, it can be observed that EDA and HR increase when performing a cognitive task. The same pattern can be observe with the EDA and ACC values as shown in Figure 3c. It is evident that EDA increases from the resting class to the cognitive class, with ACC also increasing, which indicates a positive relationship between EDA and ACC. This is understandable, as a cognitive activity can make a person adjust their position or movement, resulting in greater acceleration. Similarly, a high cognitive load can lead to an increase in the skin conductance (EDA) as also demonstrated by [61].

### 6.2 Limitations

While this study provides valuable insights into the classification of individualised physiological data, some limitations need to be noted, specifically (1) the size of the datasets, (2) the environment of the study, and (3) the simplicity of the data.

First, the study included 10 minutes worth of data for each of the resting and cognitive activities, which was further cut down after data processing to 8 minutes each (as discussed in Section 4). After aggregating the data, there were a total of 409 instances for each participant. When training and testing on all participants (all participants evaluation and leave-one-out evaluation), this results in a total dataset of 10,634 instances. However, when considering each participant alone (individual evaluation) this results in 26 datasets of only 409 instances each. This is a considerably small dataset for machine learning, which could attribute to the machine learning models high accuracy rate. With a limited amount of data, achieving an accuracy of 97% could be a result of an overfitted model rather than a robust one. For instance, the model may be memorising instances instead of actually learning data patterns during the training phase.

Second, the study was conducted in a controlled environment and real-world conditions may introduce additional factors and challenges that were not addressed in this project. Researchers often discuss the differences between laboratory-based and in-situ studies. König et al. [35], for example, discuss the challenges when conducting research on-site out in industry. While our controlled environment was satisfactory as a preliminary study, it should be noted that the collection of physiological data in real-world scenarios tends to result in significantly more complex readings. For example, our participants were sitting still during the cognitively intensive task. This allowed us to more easily identify cognitive workload using ECG and EDA. However, in a real-world scenario, participants may be moving around and completing other tasks simultaneously. This would complicate the readings that are received, e.g., heart rate may reflect both physical activity and cognitive workload.

Finally, the distribution of data as seen in Section 6.1 is uniform and there is a clear separation between data points of both classes for all types of physiological reading. While this suggests that the model may not be overfitting, it also indicates the low complexity of the data. This can lead to a lack of variability, which in turn can indicate that the data is not representative of real-world scenarios. As a result, the model might not be as robust when applied to practical settings. This is of particular relevance for the individual evaluation method, as evaluating each participant individually removes a lot of the variability in the data points.

### 6.3 Future work

There are three main areas in which we would like to expand this study. First, Section 6.1 outlined the visualisation of one participant's data. The next step would be to visualise the data for each participant, thereby identifying trends across participants and investigating whether all participant data is as segmented as our first visualisation.

Second, we would like to repeat this study with an increased size of the dataset for each participant. For example, by increasing the number of sessions from two to ten, the size of each individual dataset would increased from 409 to 2,045 instances. This would allow us to investigate the variation in physiological readings, not just between participants, but for each individual participant. Collecting resting and cognitive data from each participant on ten different occasions would allow for greater variation of days and times.

Third, we would like to extend the study with a focus on the complexity of the data. While the current study was conducted in a controlled environment, one of the next steps would be to conduct the same study in a less controlled environment. This would highlight whether real-world conditions result in less segmented and more complex patterns, and introduce additional factors and challenges that were not addressed in this project.

## 7 Conclusion

This paper investigates different ways of approaching participatory data. Using cognitive fatigue as a case study, we evaluate the effectiveness of treating participants collectively or as individuals. We demonstrate this by analyzing a dataset of physiological data from 26 participants, which included measures such as heart rate, heart rate variability, skin conductance, and movement. Our findings revealed that evaluation methods that treated each participant as an individual achieved an average accuracy of 97%, compared to 74% and 73% for methods that combined all participants data. These results highlight the potential improvements when acknowledging the unique physiological profiles of individuals. Therefore, incorporating individualized analysis into the study of wearable technology and physiological data could enhance the precision and relevance of such research.

## References


[1] 2021. https://learnetutorials.com/machine-learning/support-vector-machines

[2] 2022. https://learnetutorials.com/machine-learning/bagging-and-random-forest

[3] Usman Alhaji Abdurrahman, Shih-Ching Yeh, Yunying Wong, and Liang Wei. 2021. Effects of neuro-cognitive load on learning transfer using a virtual reality-based driving system. *Big Data and Cognitive Computing* 5, 4 (2021), 54.

[4] Ankita Agarwal, Josephine Graft, Noah Schroeder, and William Romine. 2021. Sensor-Based Prediction of Mental Effort during Learning from Physiological Data: A Longitudinal Case Study. *Signals* 2, 4 (2021), 886–901.

[5] Muneeb Imtiaz Ahmad, Ingo Keller, David A Robb, and Katrin S Lohan. 2020. A framework to estimate cognitive load using physiological data. *Personal and Ubiquitous Computing* (2020), 1–15.

[6] Ayca Aygun, Boyang Lyu, Thuan Nguyen, Zachary Haga, Shuchin Aeron, and Matthias Scheutz. 2022. Cognitive workload assessment via eye gaze and eeg in an interactive multi-modal driving task. In *Proceedings of the 2022 International Conference on Multimodal Interaction*. 337–348.

[7] Ayca Aygun, Thuan Nguyen, Zachary Haga, Shuchin Aeron, and Matthias Scheutz. 2022. Investigating methods for cognitive workload estimation for assistive robots. *Sensors* 22, 18 (2022), 6834.

[8] Patricia Besson, Christophe Bourdin, Lionel Bringoux, Erick Dousset, Christophe Maïano, Tanguy Marqueste, Daniel R Mestre, Sophie Gaetan, Jean-Pierre Baudry, and Jean-Louis Vercher. 2013. Effectiveness of physiological and psychological features to estimate helicopter pilots' workload: A Bayesian network approach. *IEEE Transactions on Intelligent Transportation Systems* 14, 4 (2013), 1872–1881.

[9] Patricia Besson, Christophe Maïano, Lionel Bringoux, Tanguy Marqueste, Daniel R Mestre, Christophe Bourdin, Erick Dousset, Mathilde Durand, and Jean-Louis Vercher. 2012. Cognitive workload and affective state: A computational study using Bayesian networks. In *2012 6th IEEE International Conference Intelligent Systems*. IEEE, 140–145.

[10] Andrea Bettoni, Elias Montini, Massimiliano Righi, Valeria Villani, Radostin Tsvetanov, Stefano Borgia, Cristian Secchi, and Emanuele Carpanzano. 2020. Mutualistic and adaptive human-machine collaboration based on machine learning in an injection moulding manufacturing line. *Procedia CIRP* 93 (2020), 395–400.

[11] Vadim Borisov, Enkelejda Kasneci, and Gjergji Kasneci. 2021. Robust cognitive load detection from wrist-band sensors. *Computers in Human Behavior Reports* 4 (2021), 100116.

[12] Judy Bowen, Annika Hinze, and Christopher Griffiths. 2019. Investigating real-time monitoring of fatigue indicators of New Zealand forestry workers. *Accident Analysis & Prevention* 126 (2019), 122–141.

[13] Tetiana Buraha, Jan Schneider, Daniele Di Mitri, and Daniel Schiffner. 2021. Analysis of the "D'oh!" Moments. Physiological Markers of Performance in Cognitive Switching Tasks. In *Technology-Enhanced Learning for a Free, Safe, and Sustainable World: 16th European Conference on Technology Enhanced Learning, EC-TEL 2021, Bolzano, Italy, September 20-24, 2021, Proceedings 16*. Springer, 137–148.

[14] Matthew S Caywood, Daniel M Roberts, Jeffrey B Colombe, Hal S Greenwald, and Monica Z Weiland. 2017. Gaussian process regression for predictive but interpretable machine learning models: An example of predicting mental workload across tasks. *Frontiers in human neuroscience* 10 (2017), 647.

[15] J Cegarra, B Valéry, Eugénie Avril, C Calmettes, and J Navarro. 2020. OpenMATB: A Multi-Attribute Task Battery promoting task customization, software extensibility and experiment replicability. *Behavior research methods* 52 (2020), 1980–1990.

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

[17] Fabio Dell'Agnola, Una Pale, Rodrigo Marino, Adriana Arza, and David Atienza. 2021. Mbiotracker: Multimodal self-aware bio-monitoring wearable system for online workload detection. *IEEE Transactions on Biomedical Circuits and Systems* 15, 5 (2021), 994–1007.

[18] Na Du, Feng Zhou, Elizabeth M Pulver, Dawn M Tilbury, Lionel P Robert, Anuj K Pradhan, and X Jessie Yang. 2020. Predicting driver takeover performance in conditionally automated driving. *Accident Analysis & Prevention* 148 (2020), 105748.

[19] Colin Elkin and Vijay Devabhaktuni. 2019. Analysis of alternatives for neural network training techniques in assessing cognitive workload. In *Advances in Neuroergonomics and Cognitive Engineering: Proceedings of the AHFE 2018 International Conference on Neuroergonomics and Cognitive Engineering, July 21–25, 2018, Loews Sapphire Falls Resort at Universal Studios, Orlando, Florida USA 9*. Springer, 27–37.

[20] Justin R Estepp and James C Christensen. 2015. Electrode replacement does not affect classification accuracy in dual-session use of a passive brain-computer interface for assessing cognitive workload. *Frontiers in Neuroscience* 9 (2015), 87526.

[21] Peyvand Ghaderyan and Ataollah Abbasi. 2021. Sparse coding classification and cepstral singular value for cognitive workload estimation. *Computers & Electrical Engineering* 91 (2021), 107031.

[22] Anmol Gupta, Ronnie Daniel, Akash Rao, Partha Pratim Roy, Sushil Chandra, and Byung-Gyu Kim. 2023. Raw electroencephalogram-based cognitive workload classification using directed and nondirected functional connectivity analysis and Deep Learning. *Big Data* 11, 4 (2023), 307–319.

[23] Anmol Gupta, Gourav Siddhad, Vishal Pandey, Partha Pratim Roy, and Byung-Gyu Kim. 2021. Subject-specific cognitive workload classification using EEG-based functional connectivity and deep learning. *Sensors* 21, 20 (2021), 6710.

[24] Dengbo He, Martina Risteska, Birsen Donmez, and Kaiyang Chen. 2021. Driver cognitive load classification based on physiological data—case study 7. In *Intelligent Computing for Interactive System Design: Statistics, Digital Signal Processing, and Machine Learning in Practice*. 409–429.

[25] Ryan G Hefron, Brett J Borghetti, James C Christensen, and Christine M Schubert Kabban. 2017. Deep long short-term memory structures model temporal dependencies improving cognitive workload estimation. *Pattern Recognition Letters* 94 (2017), 96–104.

[26] Annika Hinze, Jemma L König, and Judy Bowen. 2021. Worker-fatigue contributing to workplace incidents in New Zealand Forestry. *Journal of safety research* 79 (2021), 304–320.

[27] Niraj Hirachan, Anita Mathews, Julio Romero, and Raul Fernandez Rojas. 2022. Measuring cognitive workload using multimodal sensors. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 4921–4924.

[28] Jing Huang, Yu Liu, and Xiaoyan Peng. 2022. Recognition of driver's mental workload based on physiological signals, a comparative study. *Biomedical Signal Processing and Control* 71 (2022), 103094.

[29] Mir Riyanul Islam, Shaibal Barua, Mobyen Uddin Ahmed, Shahina Begum, Pietro Aricò, Gianluca Borghini, and Gianluca Di Flumeri. 2020. A novel mutual information based feature set for drivers' mental workload evaluation using machine learning. *Brain Sciences* 10, 8 (2020), 551.

[30] Taikyeong Jeong. 2020. Time-series data classification and analysis associated with machine learning algorithms for cognitive perception and phenomenon. *IEEE Access* 8 (2020), 222417–222428.

[31] Wonse Jo, Ruiqi Wang, Go-Eum Cha, Su Sun, Revanth Krishna Senthilkumaran, Daniel Foti, and Byung-Cheol Min. 2024. Mocas: A multimodal dataset for objective cognitive workload assessment on simultaneous tasks. *IEEE Transactions on Affective Computing* (2024).

[32] Apostolos Kalatzis, Ashish Teotia, Vishnunarayan Girishan Prabhu, and Laura Stanley. 2021. A database for cognitive workload classification using electrocardiogram and respiration signal. In *Advances in Neuroergonomics and Cognitive Engineering: Proceedings of the AHFE 2021 Virtual Conferences on Neuroergonomics and Cognitive Engineering, Industrial Cognitive Ergonomics and Engineering Psychology, and Cognitive Computing and Internet of Things, July 25-29, 2021, USA*. Springer, 509–516.

[33] Hyun Suk Kim, Daesub Yoon, Hyun Soon Shin, and Cheong Hee Park. 2018. Predicting the EEG level of a driver based on driving information. *IEEE transactions on intelligent transportation systems* 20, 4 (2018), 1215–1225.

[34] Jemma L König, Annika Hinze, and Judy Bowen. 2023. Workload categorization for hazardous industries: The semantic modelling of multi-modal physiological data. *Future Generation Computer Systems* 141 (2023), 369–381.

[35] Jemma L König, Jascha Penaredondo, Emily McCullagh, Judy Bowen, and Annika Hinze. 2023. Let's Make it Accessible: The Challenges Of Working With Low-cost Commercially Available Wearable Devices. In *Proceedings of the 35th Australian Computer-Human Interaction Conference*. 493–503.

[36] Thomas Kosch, Jakob Karolus, Johannes Zagermann, Harald Reiterer, Albrecht Schmidt, and Paweł W. Woźniak. 2023. A Survey on Measuring Cognitive Workload in Human-Computer Interaction. *ACM Comput. Surv.* 55, 13s, Article 283 (jul 2023), 39 pages.

[37] Serena Lee-Cultura, Kshitij Sharma, Sofia Papavlasopoulou, and Michail Giannakos. 2020. Motion-based educational games: Using multi-modal data to predict player's performance. In *2020 IEEE conference on games (cog)*. IEEE, 17–24.

[38] Chiuhsiang Joe Lin and Rio Prasetyo Lukodono. 2022. Classification of mental workload in Human-robot collaboration using machine learning based on physiological feedback. *Journal of Manufacturing Systems* 65 (2022), 673–685.

[39] Jesus L Lobo, Javier Del Ser, Flavia De Simone, Roberta Presta, Simona Collina, and Zdenek Moravek. 2016. Cognitive workload classification using eye-tracking and EEG data. In *Proceedings of the international conference on human-computer interaction in aerospace*. 1–8.

[40] Tiffany Luong, Nicolas Martin, Anais Raison, Ferran Argelaguet, Jean-Marc Diverrez, and Anatole Lécuyer. 2020. Towards real-time recognition of users mental workload using integrated physiological sensors into a VR HMD. In *2020 IEEE international symposium on mixed and augmented reality (ISMAR)*. IEEE, 425–437.

[41] Yue Ma, Qing Liu, and Liu Yang. 2022. Exploring seafarers' workload recognition model with EEG, ECG and task scenarios' complexity: a bridge simulation study. *Journal of Marine Science and Engineering* 10, 10 (2022), 1438.

[42] Emma MacNeil, Ashley Bishop, and Kurtulus Izzetoglu. 2022. Study of Different Classifiers and Multi-modal Sensors in Assessment of Workload. In *International Conference on Human-Computer Interaction*. Springer, 151–161.

[43] Anthony D McDonald, Thomas K Ferris, and Tyler A Wiener. 2020. Classification of driver distraction: A comprehensive analysis of feature generation, machine learning, and input measures. *Human factors* 62, 6 (2020), 1019–1035.

[44] Quentin Meteier, Marine Capallera, Simon Ruffieux, Leonardo Angelini, Omar Abou Khaled, Elena Mugellini, Marino Widmer, and Andreas Sonderegger. 2021. Classification of drivers' workload using physiological signals in conditional automation. *Frontiers in psychology* 12 (2021), 596038.

[45] Jadon Miller, Mitchell A Head, Mahonri W Owen, Merel Cornelie Janna Hoskens, Jemma L Konig, and Judy Bowen. 2023. First Do No Harm: Cultural and Ethical Impacts of User Studies. In *Proceedings of the 35th Australian Computer-Human Interaction Conference*. 71–77.

[46] Yuna Noh, Seyun Kim, Young Jae Jang, and Yoonjin Yoon. 2021. Modeling individual differences in driver workload inference using physiological data. *International journal of automotive technology* 22 (2021), 201–212.

[47] Domen Novak, Matjaž Mihelj, and Marko Munih. 2009. Using Psychophysiological Measurements in Physically Demanding Virtual Environments. In *Proceedings of the 12th IFIP TC 13 International Conference on Human-Computer Interaction: Part I* (Uppsala, Sweden) *(INTERACT '09)*. Springer-Verlag, Berlin, Heidelberg, 490–493.

[48] Hyuk Oh, Bradley D Hatfield, Kyle J Jaquess, Li-Chuan Lo, Ying Ying Tan, Michael C Prevost, Jessica M Mohler, Hartley Postlethwaite, Jeremy C Rietschel, Matthew W Miller, et al. 2015. A composite cognitive workload assessment system in pilots under various task demands using ensemble learning. In *Foundations of Augmented Cognition: 9th International Conference, AC 2015, Held as Part of HCI International 2015, Los Angeles, CA, USA, August 2–7, 2015, Proceedings 9*. Springer, 91–100.

[49] Anaïs Pontiggia, Danielle Gomez-Merino, Michael Quiquempoix, Vincent Beauchamps, Alexis Boffet, Pierre Fabries, Mounir Chennaoui, and Fabien Sauvet. 2024. MATB for assessing different mental workload levels. *Frontiers in Physiology* 15 (2024), 1408242.

[50] Yuning Qiu, Teruhisa Misu, and Carlos Busso. 2019. Analysis of the relationship between physiological signals and vehicle maneuvers during a naturalistic driving study. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, 3230–3235.

[51] Hamidur Rahman, Mobyen Uddin Ahmed, Shaibal Barua, and Shahina Begum. 2020. Non-contact-based driver's cognitive load classification using physiological and vehicular parameters. *Biomedical Signal Processing and Control* 55 (2020), 101634.

[52] William L Romine, Noah L Schroeder, Josephine Graft, Fan Yang, Reza Sadeghi, Mahdieh Zabihimayvan, Dipesh Kadariya, and Tanvi Banerjee. 2020. Using machine learning to train a wearable device for measuring students' cognitive load during problem-solving activities based on electrodermal activity, body temperature, and heart rate: Development of a cognitive load tracker for both personal and classroom use. *Sensors* 20, 17 (2020), 4833.

[53] Raphaëlle N Roy, Nicolas Drougard, Thibault Gateau, Frédéric Dehais, and Caroline PC Chanel. 2020. How can physiological computing benefit human-robot interaction? *Robotics* 9, 4 (2020), 100.

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

[55] Harshita Sharma, Lior Drukker, Aris T Papageorghiou, and J Alison Noble. 2021. Machine learning-based analysis of operator pupillary response to assess cognitive workload in clinical ultrasound imaging. *Computers in biology and medicine* 135 (2021), 104589.

[56] Kshitij Sharma, Evangelos Niforatos, Michail Giannakos, and Vassilis Kostakos. 2020. Assessing cognitive performance using physiological and facial features: Generalizing across contexts. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020), 1–41.

[57] Mansi Sharma and Ela Kumar. 2022. A Review on Estimation of Workload from Electroencephalogram (EEG) Using Machine Learning. In *International Conference on Advancements in Interdisciplinary Research*. Springer, 255–264.

[58] Erin T Solovey, Marin Zec, Enrique Abdon Garcia Perez, Bryan Reimer, and Bruce Mehler. 2014. Classifying driver workload using physiological and driving performance data: two field studies. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 4057–4066.

[59] Alexis D Souchet, Mamadou Lamarana Diallo, and Domitile Lourdeaux. 2022. Cognitive load classification with a stroop task in virtual reality based on physiological data. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 656–666.

[60] Hamed Taheri Gorji, Nicholas Wilson, Jessica VanBree, Bradley Hoffmann, Thomas Petros, and Kouhyar Tavakolian. 2023. Using machine learning methods and EEG to discriminate aircraft pilot cognitive workload during flight. *Scientific Reports* 13, 1 (2023), 2507.

[61] Peter Thorvald and Jessica Lindblom. 2014. Initial development of a cognitive load assessment tool. *Advances in cognitive engineering and neuroergonomics* (2014), 223–232.

[62] Harshita Ved and Caglar Yildirim. 2021. Detecting mental workload in virtual reality using eeg spectral data: A deep learning approach. In *2021 IEEE international conference on artificial intelligence and virtual reality (AIVR)*. IEEE, 173–178.

[63] Nicholas Wilson, Hamed Taheri Gorji, Jessica VanBree, Bradley Hoffmann, Kouhyar Tavakolian, and Thomas Petros. 2021. Identifying opportunities for augmented cognition during live flight scenario: an analysis of pilot mental workload using EEG. In *94th International Symposium on Aviation Psychology*. 444.

[64] Haochen Wu, Charne C Folks, Alparslan Emrah Bayrak, Jonathon M Semereka, Bogdan I Epureanu, US Army Aberdeen Test Center, and MD Aberdeen. 2022. Human-Autonomy Teaming in Immersive Environments. (2022).

[65] Grace C Wusk, Andrew F Abercromby, and Hampton C Gabler. 2019. Psychophysiological monitoring of aerospace crew state. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*. 404–407.

[66] Shuo Yang, Zhong Yin, Yagang Wang, Wei Zhang, Yongxiong Wang, and Jianhua Zhang. 2019. Assessing cognitive mental workload via EEG signals and an ensemble deep learning classifier based on denoising autoencoders. *Computers in biology and medicine* 109 (2019), 159–170.

[67] Cho Yin Yiu, Kam KH Ng, Xinyu Li, Xiaoge Zhang, Qinbiao Li, Hok Sam Lam, and Man Ho Chong. 2022. Towards safe and collaborative aerodrome operations: Assessing shared situational awareness for adverse weather detection with EEG-enabled Bayesian neural networks. *Advanced Engineering Informatics* 53 (2022), 101698.

[68] Renato Zanetti, Adriana Arza, Amir Aminifar, and David Atienza. 2021. Real-time EEG-based cognitive workload monitoring on wearable devices. *IEEE transactions on biomedical engineering* 69, 1 (2021), 265–277.

[69] Xiaoge Zhang, Sankaran Mahadevan, Nathan Lau, and Matthew B Weinger. 2020. Multi-source information fusion to assess control room operator performance. *Reliability Engineering & System Safety* 194 (2020), 106287.

[70] Guozhen Zhao, Yong-Jin Liu, and Yuanchun Shi. 2018. Real-time assessment of the cross-task mental workload using physiological measures during anomaly detection. *IEEE Transactions on Human-Machine Systems* 48, 2 (2018), 149–160.

[71] Xin Zhao. 2018. *Driver Cognitive Workload Detection via Eye-tracking and Physiological Modalities*. University of Toronto (Canada).

[72] Zhanpeng Zheng, Zhong Yin, Yongxiong Wang, and Jianhua Zhang. 2023. Inter-subject cognitive workload estimation based on a cascade ensemble of multilayer autoencoders. *Expert Systems with Applications* 211 (2023), 118694.

[73] Zhanpeng Zheng, Zhong Yin, and Jianhua Zhang. 2020. An ELM-based Deep SDAE Ensemble for Inter-Subject Cognitive Workload Estimation with Physiological Signals. In *2020 39th Chinese Control Conference (CCC)*. IEEE, 6237–6242.

Received August 2024