Article

Decoding Mental Effort in a Quasi-Realistic Scenario: A Feasibility Study on Multimodal Data Fusion and Classification

Sabrina Gado [1,†], Katharina Lingelbach [2,3,†,*], Maria Wirzberger [4,5,‡] and Mathias Vukelić [2,‡]

[1] Experimental Clinical Psychology, Department of Psychology, Julius-Maximilians-University of Würzburg, Würzburg, Germany
[2] Applied Neurocognitive Systems, Fraunhofer Institute for Industrial Engineering IAO, Stuttgart, Germany
[3] Applied Neurocognitive Psychology Lab, Department of Psychology, Carl von Ossietzky University, Oldenburg, Germany
[4] University of Stuttgart, Department of Teaching and Learning with Intelligent Systems, Stuttgart, Germany
[5] LEAD Graduate School & Research Network, University of Tuebingen, Tuebingen, Germany
* Correspondence: Katharina.Lingelbach@iao.fraunhofer.de
† These authors contributed equally to this work and share first authorship.
‡ These authors share last authorship.

Abstract: Humans' performance varies due to the mental resources that are available to successfully pursue a task. To monitor users' current cognitive resources in naturalistic scenarios, it is essential to not only measure demands induced by the task itself but also consider situational and environmental influences. We conducted a multimodal study with 18 participants (nine female, M = 25.9 with SD = 3.8 years). In this study, we recorded, respiratory, ocular, cardiac, and brain activity using functional near-infrared spectroscopy (fNIRS) while participants performed an adapted version of the warship commander task with concurrent emotional speech distraction. We tested the feasibility of decoding the experienced mental effort with a multimodal machine learning architecture. The architecture comprised feature engineering, model optimisation, and model selection to combine multimodal measurements in a cross-subject classification. Our approach reduces possible overfitting and reliably distinguishes two different levels of mental effort. These findings contribute to the prediction of different states of mental effort and pave the way toward generalised state monitoring across individuals in realistic applications..

Keywords: mental effort; machine learning; multimodal physiological signals; sensor fusion; neuroergonomics; human-machine interaction

1. Introduction

In everyday life, we constantly face situations demanding high stakes for maximum gains; for instance, to succeed in rapidly acquiring complex cognitive skills or making decisions under high pressure. Thereby, a fit between personal skills and the task's requirements determines the quality of outcomes. This fit is vital, especially in performance-oriented contexts such as learning and training, safety-critical monitoring, or high-risk decision-making. A person's performance can be affected by several factors: 1) level of experience and skills, 2) current physical conditions (e.g., illness or fatigue), 3) current psychological conditions (e.g., stress, motivation, or emotions), or 4) external circumstances (e.g., noise, temperature, or distractions; Hart and Staveland [1], Young *et al.* [2]).

To reliably quantify the mental effort during a particular task, different measures can be used: 1) behavioural (i.e., performance-based), 2) subjective, and 3) neurophysiological measures [3–5]. While performance can be inspected by tracking the user's task-related progress, the actual pattern of invested cognitive resources can only be derived by measuring brain activity with neuroimaging techniques. Coupled with sophisticated signal processing and machine learning, advances in portable neuroimaging techniques have paved the way for studying mental effort and its possible influences

 doi:10.20944/preprints202306.1083.v1

from a neuroergonomic perspective [6,7]. Recently, functional near-infrared spectroscopy (fNIRS) has been used to study cognitive and emotional processes with high ecological validity [6,8–10]. fNIRS is an optical imaging technology allowing researchers to measure local oxy-haemoglobin (HbO) and deoxy-haemoglobin (HbR) changes in cortical regions. Higher mental effort is associated with an increase of HbO and a decrease of HbR in the prefrontal cortex (PFC) [11–13]. The PFC is crucial for executive functions like maintaining goal-directed behaviour and suppressing goal-irrelevant distractions [14,15]. In addition to changes in the central nervous system, an increased mental effort also leads to changes in the autonomic nervous system. The autonomic nervous system, as part of the peripheral nervous system, regulates automatic physiological processes to maintain homeostasis in bodily functioning [16,17]. Increased mental effort is associated with decreased parasympathetic nervous system activity and increased sympathetic nervous system activity [18–20]. Typical correlates of the autonomic nervous system for cognitive demands, engagement or mental effort are cardiac activity (e.g., heart rate and heart rate variability), respiration (rate, airflow, and volume), electrodermal activity (skin conductance level and response), blood pressure, body temperature, and ocular measures like pupil dilation, blinks, and eye movements [7,19,21–24].

Not surprisingly, all these measures are, thus, often used as a stand-alone indicator for mental effort (i.e., in a *unimodal approach*). However, a *multimodal approach* has several advantages over using only one measure. It can compensate for specific weaknesses and profit from the strengths of the different complementary measurement methods (performance, subjective experience as well as neuro- and peripheral physiological measures) [25–27]. For instance, (neuro-)physiological measures can be obtained without imposing an additional task [16] and allow for capturing cognitive subprocesses involved in executing the primary task [28]. A multimodal approach, hence, provides a more comprehensive view of (neuro-)physiological processes related to mental effort [4,5,25,29], as it can capture both central and peripheral nervous system processes [21,27]. However, fusing data from different sources remains a major challenge for multimodal approaches. Machine learning (ML) methods provide solutions to compare and combine data streams from different measurements. ML algorithms are becoming increasingly popular in computational neuroscience [30,31]. The rationale behind these algorithms is that the relationship between several input data streams and a particular outcome variable, e.g., mental effort, can be estimated from the data by iteratively fitting and adapting the respective models. This allows for data-driven analyses and provides ways to exploratorily identify patterns in the data that are informative [32].

Data-driven approaches can also be advantageous in bridging the disparity between laboratory research and real-world applications. For instance, when specific temporal events (such as a stimulus onset) or the brain correlates of interest, are not precisely known. In contrast to traditional laboratory studies that typically rely on simplified and artificial stimuli and tasks, a naturalistic approach seeks to emulate, to some extent, the intricacy of real-world situations. Hence, these studies can provide insights into how the brain processes information and responds to complex stimuli in the real world [33].

Real-world settings are usually characterised by multiple situational characteristics including concurrent distractions that affect the allocation of attentional and cognitive resources [34]. According to the working memory model by Baddeley and Hitch [35], performance is notably diminished when distractions deplete resources from the same modality as the primary task. However, Soerqvist *et al.* [36] propose the involvement of cognitive control mechanisms that result in reduced processing of task-irrelevant information under higher mental effort. To uphold task-relevant cognitive processes, high-level cortical areas, particularly the prefrontal cortex (PFC), which govern top-down regulation and executive functioning, suppress task- or stimulus-irrelevant neural activities by inhibiting the processing of distractions [28]. Consequently, the effects of distractors are mitigated. In light of these considerations, understanding the capacity of a stimulus to capture attention in a bottom-up manner, known as salience, emerges as a crucial aspect. A salient stimulus has the potential to disrupt top-down goal-oriented and intentional attention processes [37] and to impair performance in a primary task

 doi:10.20944/preprints202306.1083.v1

[38–40]. Previous studies found that irrelevant, yet intelligible speech exerts such disruptive effects on participants' performance in complex cognitive tasks [41,42]. Consequently, intelligible speech might heighten the salience of a distracting stimulus. Moreover, further studies revealed that the emotional intensity and valence of a stimulus also play a role in influencing its salience [37,43]. Despite their detrimental impact on performance, people frequently experience such salient distractions (such as verbal utterances from colleagues) at work, even in highly demanding safety-relevant tasks. Therefore, gaining an understanding of the underlying cognitive processes in naturalistic scenarios and identifying critical moments that lead to performance decreases in real-world settings are crucial research topics in the field of neuroergonomics.

To decode and predict cognitive states, most research so far focused on subject-dependent classification. These approaches face the challenge of high inter-individual variability in physiological signals when generalising the model to others [44]. Recently, pioneering efforts have been made to develop cross-subject models that overcome the need for subject-specific information during training [45,46]. Solutions to address the challenge of inter-individual variability [47] are crucial for the development of "plug and play" real-time state recognition systems [48] as well as the resource-conserving exploitation of already available large datasets without time-consuming individual calibration sessions. Taking into account the aforementioned considerations, we present a feasibility study to decode mental effort from multimodal physiological and behavioural signals in a quasi-realistic scenario. We used an adapted warship commander task, that induces mental effort based on a combination of attentional and cognitive processes, such as object perception, object discrimination, rule application, and decision-making [49]. To create a complex close-to-naturalistic scenario, three emotional types of auditory speech-based stimuli with neutrally, positively, and negatively connotated prosody were presented during the task as concurrent distractions [50]. Concurrently, both brain-related as well as peripheral physiological signals associated with mental effort were recorded.

We hypothesised that a well-designed multimodal voting ML architecture is preferable compared to a classifier based on a) only one modality (unimodal approach) and b) a combined, unbalanced feature set of all modalities. We expected that a multimodal voting ML model is capable of predicting subjectively experienced mental effort induced by the task itself but also by the suppression of situational auditory distractions in a complex close-to-realistic environment. Thus, we first investigated whether a combined prediction of various ML models is superior to the prediction of a single model (RQ1) and, second, we explored whether a multimodal classification that combines and prioritizes the predictions of different modalities is superior to a unimodal prediction (RQ2).

2. Materials and Methods

2.1. Participants

Interested volunteers filled in a screening questionnaire that checked eligibility for study participation and collected demographic characteristics. Individuals with insufficient knowledge of the German language or limited colour vision were excluded due to a lack of ability to perform the task. Further, we did not include pregnant women, participants indicating precarious alcohol or any drug consumption, as well as those reporting mental, neurological, or cardiovascular diseases. Another exclusion criterion was the presence of an implant or surgery in the head area. Because data were collected in June 2021 during the COVID-19 pandemic, we refrained from inviting any participants to the laboratory who belong to the risk group for severe COVID-19 disease, according to the Robert Koch Institute. The sample consisted of 18 participants (nine female, three left-handed, mean age of 25.9 years, $SD = 3.8$, range = 21–35 years). All participants had normal or corrected-to-normal vision. Before their participation, they signed an informed consent according to the recommendations of the Declaration of Helsinki and received monetary compensation for their voluntary participation. The

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2023 doi:10.20944/preprints202306.1083.v1

study was approved by the ethics committee of the Medical Faculty of the University of Tuebingen, Germany (*ID* : 827/2020*BO*1).

2.2. Experimental Task

Participants performed an adapted version of a warship commander task (WCT [51]; adapted by Becker *et al.* [49]). The WCT is a quasi-realistic navy command and control task designed as a basic analogue to a Navy air warfare task [52]. It is suitable to investigate various cognitive processes of human decision-making and action execution [52]. Here, we used a non-military safety-critical task, where participants had to identify two different flying objects on a simulated radar screen around an airport. Objects included either registered drones (neutral, non-critical objects), or non-registered (critical) drones. They had to prevent the non-registered drones, potentially being a safety issue, from entering the airport's air space. Non-registered drones entering pre-defined ranges close to the airport had to be first warned and then repelled in the next step. A performance score was computed based on participants' accuracy and reaction time. See Becker *et al.* [49], for a more detailed description of the scoring system and see Figure 1 for an overview of the interface.

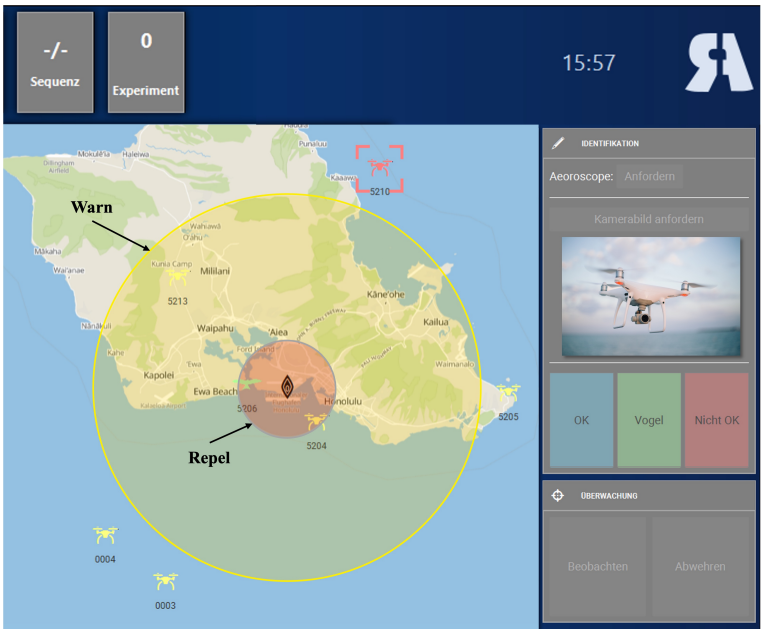

Figure 1. Elements of the WCT interface. Left side of the screen (map): Participants had to monitor the aerial space of the airport. When an unregistered drone entered the yellow area (outer circle), participants had to warn that drone; when an unregistered drone entered the red area (inner circle), participants had to repel it. Right side of the screen (graphical user interface): Participants had to request codes and pictures of unknown flying objects and then classify them as birds, registered drones, or unregistered drones.

During the task, we presented vocal utterances, either spoken in a happy, angry, or neutral way from the Berlin Database of Emotional Speech (Emo-DB [50]). These utterances were combined into different audio files, each one minute long, with speakers and phrases randomly selected and as little repetition as possible within each file. We also included a control condition where no auditory distraction was presented. The task load was manipulated by implementing two difficulty levels in the WCT (low and high). This resulted in a 2 × 4 design with eight experimental conditions. Participants completed two rounds of all conditions in the experiment. Before the respective round, a resting state measurement was conducted (30 seconds). Each round then consisted of eight blocks each comprising three 60-second trials of the same experimental condition. The task load condition (operationalised with the difficulty level) was alternated across blocks. Half of the subjects started with a high task load and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2023 doi:10.20944/preprints202306.1083.v1

the other half with a low task load block. Similarly, the concurrent emotional condition (operationalised with different auditory distractions) was randomised and sampled without replacement. Before each block, except for the first, participants completed a baseline condition trial with a very low difficulty level where they had to track six objects, of which three were non-registered drones. In the low task load condition, participants had to track 12 objects, of which six were non-registered drones. In the high task load condition, they had to track 36 objects, of which 17 were non-registered drones. We used different emotional audio files for the trials in one block. Before and after the whole experiment, as well as after each experimental block, participants filled in questionnaires. See Figure 2 for a schematic representation of the whole experimental procedure. Overall, the experiment lasted approximately 120 minutes including 30 minutes of preparation time of the used measurement devices and calibration procedures.

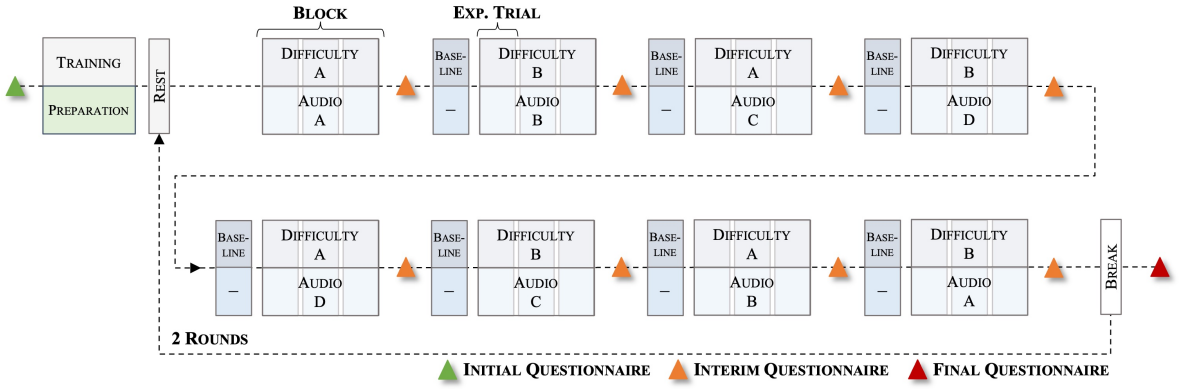

Figure 2. Procedure of the experiment. The presented procedure is exemplary as the task load condition was alternating, and the concurrent emotional condition was pseudo-randomised throughout the different blocks.

2.3. Data Collection

2.3.1. Questionnaires

Subjectively perceived mental effort and affective states were assessed after each experimental block. We used the NASA TLX effort and frustration subscales [1], EmojiGrid [53], and categorical Circumplex Affect Assessment Tool (CAAT [54]). After the experiment, participants answered questionnaires regarding personal traits that might have influenced their performance and behaviour during the study. These questionnaires comprised the short version of the German Big Five Inventory (BFI-K [55]), the German State-Trait-Anxiety Inventory (STAI [56]), the Attention and Performance Self-Assessment (APSA [57]) and the German language version of the Barratt Impulsiveness Scale-11 (BIS [58]). Here, we only used the NASA TLX ratings of mental effort for labelling the (neuro-)physiological data in the ML classification. The other subjective measures were not of interest in this analysis.

2.3.2. Eye-Tacking, Physiology, and Brain Activity

The ocular activity was recorded with the screen-based Tobii Pro Spectrum eye-tracking system, which provides gaze position and pupil dilation data at a sampling rate of 60 Hz. To capture changes in physiological responses, participants were wearing a BioHarness™ belt recording electrocardiographic (ECG), respiration, and temperature signals at a sampling rate of 1 Hz. Here, we used automatically computed, aggregated scores for the heart rate, heart rate variability, and respiration rate and amplitude from the device. Physiological as well as behavioural measures were recorded using the iMotions Biometric Research Platform software. Participants' brain activity was recorded with a NIRx NIRSport2 system which emits light at two wavelengths, 760 and 850 nm. Data was collected with the Aurora

 doi:10.20944/preprints202306.1083.v1

fNIRS recording software at a sampling rate of 5.8 Hz. To capture regions associated with mental effort, 14 source optodes and 14 detector optodes were placed over the prefrontal cortex [12,59] using the fNIRS Optodes' Location Decider (fOLD) toolbox [60] (Figure 3, for the montage). Event triggers from the experimental task were sent to iMotions and Aurora using TCP protocols and Lab Streaming Layer (LSL). Signals from the different recording and presentation systems were temporally aligned offline after the data collection.

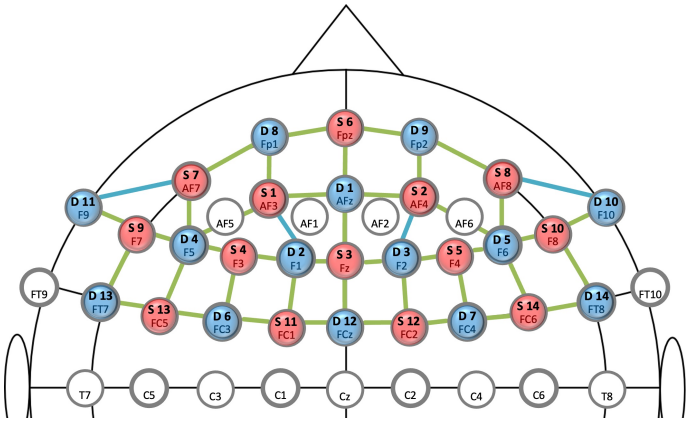

Figure 3. fNIRS optodes' location. Montage of optodes on fNIRS cap on a standard 10-20 EEG system, red optodes: sources, blue optodes: detectors, green lines: long channels, blue lines: short channels. Setup with 41 (source-detector-pairs) × 2 (wavelengths) = 82 optical channels of interest.

2.4. Data Preprocessing and Machine Learning

Data preprocessing and machine learning analyses were performed with custom-written scripts in R and python™. Continuous raw data streams were cut into non-overlapping 60-second intervals starting at the onset of each experimental trial (Figure 2). Before feeding the data into the classification pipeline, we applied the following data cleaning and preprocessing steps per modality.

2.4.1. Preprocessing of Eye-Tracking Data

Continuous eye tracker data were preprocessed using the eyetrackingR package in R [61]. Missing values were linearly interpolated and 855 trials with a length of 60 seconds (on average 47.5 trials per subject, $SD = 0.9$) were extracted. Next, we used the validity index to remove non-consistent data segments from further analysis. The index is provided by the eye tracker and indicates samples in which the eye tracker did not recognize both pupils correctly ("track loss"). 17 trials (1.99%) with a track loss proportion greater than 25% were removed, and 838 trials were left to extract fixations and pupil dilation (on average 46.6 trials per subject, $SD = 2.4$). For the preprocessing of the pupil dilation data, we used the PupillometryR R-package [62]. First, we calculated a simple linear regression of one pupil against the other and vice versa, per subject and trial to smooth out small artefacts [63]. Afterwards, we computed the mean of both pupils and filtered the data using the median of a rolling window with a size of 11 samples. To control for the variance of pupil sizes between participants, we applied a subject-wise z-score normalisation of pupil dilation. For the computation of fixations, we used the saccades R-package [64]. We obtained fixations for 565 trials (on average 31.4 trials per subject, $SD = 1.2$). To control for the variance between participants, we also computed z-scores of the number and the duration of fixations separately for each subject.

2.4.2. Preprocessing of Physiological Data

Epoching in non-overlapping 60-second time windows from the BioHarness™ raw data resulted in 832 trials (on average 46.2 trials per subject, $SD = 1.3$). We applied a correction for the between-participant variance identical to the one described for the eye-tracking data using z-score normalisation.

 doi:10.20944/preprints202306.1083.v1

2.4.3. Preprocessing of fNIRS Data

We used the libraries MNE-Python [65] and its extension MNE-NIRS [66] and guidelines from Yücel *et al.* [67] to preprocess the fNIRS data. First, we converted the raw data into an optical density measure. A channel pruning was applied using the scalp-coupling index for each channel which is an indicator of the quality of the connection between the optodes and the scalp and looks for the presence of a prominent synchronous signal in the frequency range of cardiac signals across the photo-detected signals [68]. Channels with a scalp-coupling index below 0.5 were marked as bad channels. We further applied a temporal derivative distribution repair accounting for a baseline shift and spike artefacts [69]. Channels marked as bad were interpolated, with the nearest channel providing good data quality. Afterwards, a short-separation regression was used, subtracting short-channel data from the standard long-channel signal to correct for systemic signals contaminating the brain activity measured in the long-channel [67,70]. Next, the modified Beer-Lambert Law was applied to transform optical density into oxygenated (HbO) and deoxygenated (HbR) haemoglobin concentration changes [71] with a partial pathlength factor of 6 [65]. Data were filtered using a fourth-order zero-phase Butterworth bandpass filter to remove instrumental and physiological noise (such as heartbeat and respiration; cutoff frequencies: 0.05 and 0.7 Hz; transition bandwidth: 0.02 and 0.2 Hz). HbO and HbR data was cut into epochs with a length of 60 seconds and channel-wise z-scored normalized. In total, 730 trials were obtained for the analysis (on average 40.6 trials per subject, $SD = 9.6$).

2.4.4. Feature Extraction

Our feature space comprised brain activity, physiological, ocular and performance-related measures. Table 1 gives an overview of the included features per subject and trial for each modality. We extracted the features of the fNIRS data using the mne-features package [72].

Figures 24 in the supplementary material provide exploratory analyses of the distribution and relationship between behavioural, heart activity, respiration, ocular measures, and the NASA TLX questionnaire scale effort during low and high subjective load. The Supplementary Figures 25 and 26 compare the grand average of the behavioural and physiological measures as well as single fNIRS channels of the prefrontal cortex using bootstrapping with 5000 iterations and 95% confidence intervals (CI) during low and high subjective load.

Table 1. Included Features per Modality

Modality	Features
Brain Activity	Mean, standard deviation, peak-to-peak (PTP) amplitude, skewness, and kurtosis of the 82 optical channels
Physiology	
Heart Rate	Mean, standard deviation, skewness, and kurtosis of heart rate
	Mean, standard deviation, skewness, and kurtosis of heart rate variability
Respiration	Mean, standard deviation, skewness, and kurtosis of respiration rate
	Mean, standard deviation, skewness, and kurtosis of respiration amplitude
Temperature	Mean, standard deviation, skewness, and kurtosis of body temperature
Ocular Measures	
Fixations	Number of fixations, total duration and average duration of fixations, and standard deviation of the duration of fixations
Pupillometry	Mean, standard deviation, skewness, and kurtosis of pupil dilation
Performance	Average reaction time and cumulative accuracy

2.4.5. Ground Truth for Machine Learning

Our main goal was to predict the mental effort experienced by an individual using machine learning and training data from other subjects (e.g., [73,74]). Since the experimentally manipulated task load was further influenced by situational demands (e.g., inhibiting task-irrelevant auditory emotional distraction), the perceived mental effort might not be fully captured by the experimental condition.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2023 doi:10.20944/preprints202306.1083.v1

Therefore, we explored two approaches to operationalise mental effort as a two-class classification problem: First, based on self-reports using the NASA TLX effort subscale, and second, based on the experimental task load condition.

For the mental effort prediction based on subjective perception, we performed a subject-wise median split and categorised values above the threshold as "high mental effort" and below as "low mental effort". Across all subjects, we had a mean median-based threshold of 3.8 ($SD = 3.2$, scale range = 0–20) leading to an average of 23.8 trials per subject with low mental effort ($SD = 6.6$, range = 12–39) and 14.5 trials per subject with high mental effort ($SD = 6.2$, range = 3–21; see Supplementary Figure 1 for a subject-wise distribution of the classes).

In addition, we performed a subject-wise split at the upper quartile of the NASA TLX effort subscale. The upper (or third) quartile is the point below which 75% of the data lies. We introduced this data split to also investigate the prediction and informative features of extremely high perceived mental effort, which may indicate cognitive overload. By performing a quartile split, we had a mean threshold of 6.1 ($SD = 4.3$, scale range = 0–20) across all relevant subjects (excluding subjects 5 and 9 which did not show enough variation to identify these two classes) with an average number of 30.9 low mental effort trials per subject ($SD = 8.0$, range = 16–39) and 6.6 high mental effort trials per subject ($SD = 2.3$, range = 3–9; see Supplementary Figure 2 for a subject-wise distribution of the classes).

At last, we compared the prediction of subjectively perceived mental effort with a prediction of the mental effort induced by the task, that is the experimental condition ("high task load" vs. "low task load"; see Supplementary Figure 15 for a subject-wise comparison of perceived mental effort dependent on the experimental load condition). The comparison allows to further control for confounding effects that are typical for self-reports, e.g., consistency effects or social desirability effects.

2.4.6. Model Evaluation

We fitted six machine learning approaches: 1) Logistic Regression (LR), 2) Linear Discriminant Analysis (LDA), 3) Gaussian Naïve Bayes Classifier (GNB), 4) K-Nearest Neighbor Classifier (KNN), 5) Random Forest Classifier (RFC), and 6) Support Vector Machine (SVM). They were implemented using the scikit-learn package (version 1.0.1; [75]). Figure 4 shows a schematic representation of our multimodal classification scheme and cross-subject validation procedure using multiple randomised grid search operations.

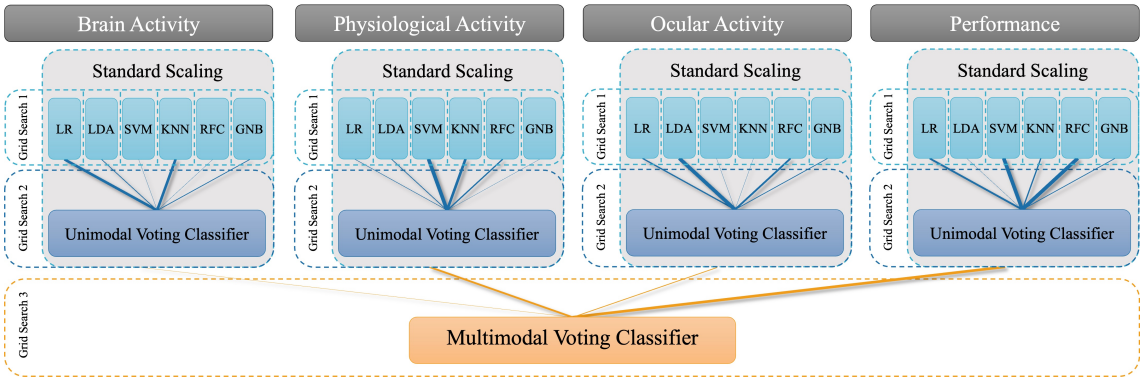

Figure 4. Classification procedure with cross-validated randomised grid searches (maximum number of 100 iterations) and a validation set consisting of one or two subjects. The first grid search optimises the hyperparameters for the different individual and unimodal classifiers. The second grid search optimises the weights as well as voting procedure (soft or hard) for the unimodal voting classifier. The third grid search optimises the weights as well as the voting procedure (soft or hard) for the multimodal voting classifier.

For the cross-subject classification, we used a leave-one-out (LOO) approach where each subject served as a test subject once (leading to 18 "outer" folds). With this 18-fold cross-subject approach,

 doi:10.20944/preprints202306.1083.v1

we simulate a scenario where a possible future system can predict an operator's current mental effort during a task without having seen any data (e.g., collected in a calibration phase) from this person before. This has the advantage that the model learns to generalise across individuals and allows the exploitation of already collected datasets from a similar context as training sets.

Our multidimensional feature space consisted of four modalities: 1) brain activity, 2) physiological activity, 3) ocular measures, and 4) performance measures. All features were z-standardised (Figure 4). This scaling ensured, that for each feature the mean is zero and the standard deviation is one, thereby, bringing all features to the same magnitude. We then trained the six classifiers (LR, LDA, GNB, KNN, RFC, and SVM) separately for each modality. Hyperparameters for each classifier were optimised by means of a cross-validated randomised grid search with a maximum number of 100 iterations and a validation set consisting of either one or two subjects. We tested both sizes of the validation set to find an optimal compromise between the robustness of the model and the required computing power. While cross-validation with two subjects counteracts the problem that the models highly adapt to an individual's unique characteristics, cross-validation with only one subject leads to a lower number of necessary iterations and a computationally more efficient approach. Due to our cross-subject approach, the selected hyperparameters varied for each predicted test subject.

Afterwards, we combined these classifiers using a voting classifier implemented in the mlxtend package (version 0.19.0 [76]). The ensemble classifier makes predictions based on aggregating the predictions of the previously trained classifiers by assigning weights to each of them. Here, we are interested in whether an ensemble approach achieves higher prediction accuracy than the best individual classifier in the ensemble. An ensemble approach has the advantage that, even if each classifier is a weak learner (meaning it does only slightly better than random prediction), the ensemble could still be a strong learner (achieving high accuracy). The voting either follows a "soft" or a "hard" voting strategy. While hard voting is based on a majority vote combining the predicted classes, soft voting considers the predicted probabilities and selects the class with the highest probability across all classifiers. The weights, as well as the voting procedure (soft or hard voting), were optimised using a third cross-validated randomised grid search with a maximum number of 100 iterations. We restricted the weights to a maximum value of 2 (range = 0–2).

With this procedure, we were able to compare the predictions of the single unimodal classifiers to a weighted combination of all classifiers of one modality.

For the multimodal approach, the voting predictions of each modality were combined into a final multimodal prediction of mental effort using a second voting classifier. This second voting classifier also assigned weights to the different modality-specific classifiers and was optimized in the same manner as the unimodal approach. We report the average F_1 score and a confusion matrix of the training set and the test subject to evaluate model performance. The F_1 score can be interpreted as a weighted average or "harmonic mean" of precision and recall (1 - good to 0 - bad performance). Precision refers to the number of samples predicted as positive that are positive (true positives). Recall measures how many of the actual positive samples are captured by the positive predictions (also called sensitivity). The F_1 score balances both aspects – identifying all positive, i.e., "high mental effort" cases, but also minimising false positives. To compare the classification performance of different models, we calculated the bootstrapped mean and its confidence intervals (CIs) over cross-validation folds with 5000 iterations per classification model. Significant differences can be derived from non-overlapping notches of the respective boxes, which mark the upper and lower boundaries of bootstrapped 95% confidence interval (CI) of the mean F_1 score. The upper CI limit of a dummy classifier represents an empirical chance level estimate (dashed grey line in all subplots of Figure 5). A dummy classifier considers only the distribution of the outcome classes for its prediction. For a prediction to be better than chance (at a significance level below .05), its bootstrapped mean must not overlap with this grey line [77]. For a significance level below .01, the lower boundaries of the CI can be used [77].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2023 doi:10.20944/preprints202306.1083.v1

3. Results

We compare the results for a mental effort prediction based on a subject-wise (1a) median and (1b) upper quartile split of the Nasa TLX effort scale as well as based on the (2) experimentally induced task load. Further, we compare two sizes of the validation set (one subject and two subjects).

3.1. Unimodal Predictions

Performance of the different modalities and classifiers is visualised in Figure 5). We do not see substantially better performance when using a larger validation set of two subjects, neither for the median split (compare Figure 5) and Supplementary Figure 3) nor for the upper quartile split (compare Supplementary Figures 7, and 11) or the prediction of the experimentally induced task load (Supplementary Figures 16 and 20). We will, therefore, focus on the models fitted with a validation set of one subject, as this is more time- and resource-efficient.

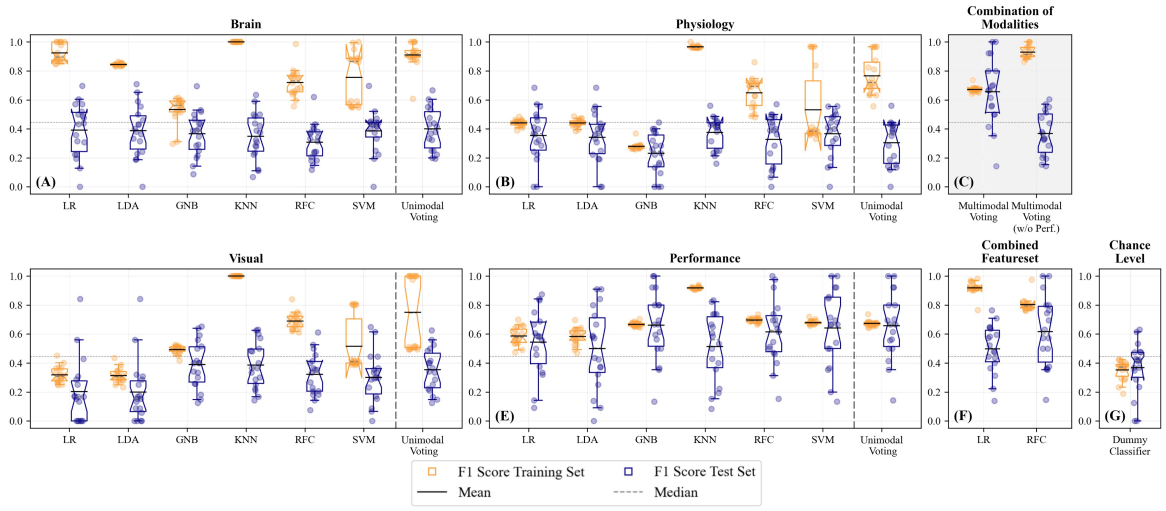

Figure 5. Prediction of the subjectively perceived mental effort based on a median split; validation set: $N = 1$. Bootstrapped 95% confidence intervals (CI; 5000 iterations) of the mean F_1 scores for the training set (left, orange) and the test set (right, blue) of the different unimodal and multimodal models. Notches in the boxes of the plot visualize the upper and lower boundary of the CI with the solid line representing the mean and the dashed grey line representing the median. The box comprises 50% of the distribution from the 25^{th} to the 75^{th} quartile. The ends of the whiskers represent the 5^{th} and 95^{th} quartile of the distribution. The continuous grey dashed line shows the upper boundary of the CI of the dummy classifier.

Figure 5 shows the performance in a median-split-based unimodal (Figure 5A, B, D, E) as well as the multimodal approach (Figure 5C; elaborated on in Section Multimodal Predictions). Regarding the unimodal classifications, we see the highest predictions of the subjectively perceived mental effort for performance data (Figure 5E compared with ocular, physiological, or brain activity measures; Figure 5A, B, and D). Except for the performance-based model, we observe overfitting indicated by the large deviation between training and test performance (Figure 5A, B, D). None of the brain activity-based models performs significantly better than the dummy classifier (Figure 5A and G) in the test data set. When examining the single classification models within each modality, the KNN, RFC, and SVM were more likely to be overfitted, as seen by the good performance in the training set but a significantly worse performance for the test subject. We combined the different classifiers using a voting classifier, of which we ascertained the voting procedure (soft vs. hard voting) and the weights with a randomised grid search. See Figure 6 for an overview of the selected voting procedures and the allocated weights per modality.

 doi:10.20944/preprints202306.1083.v1

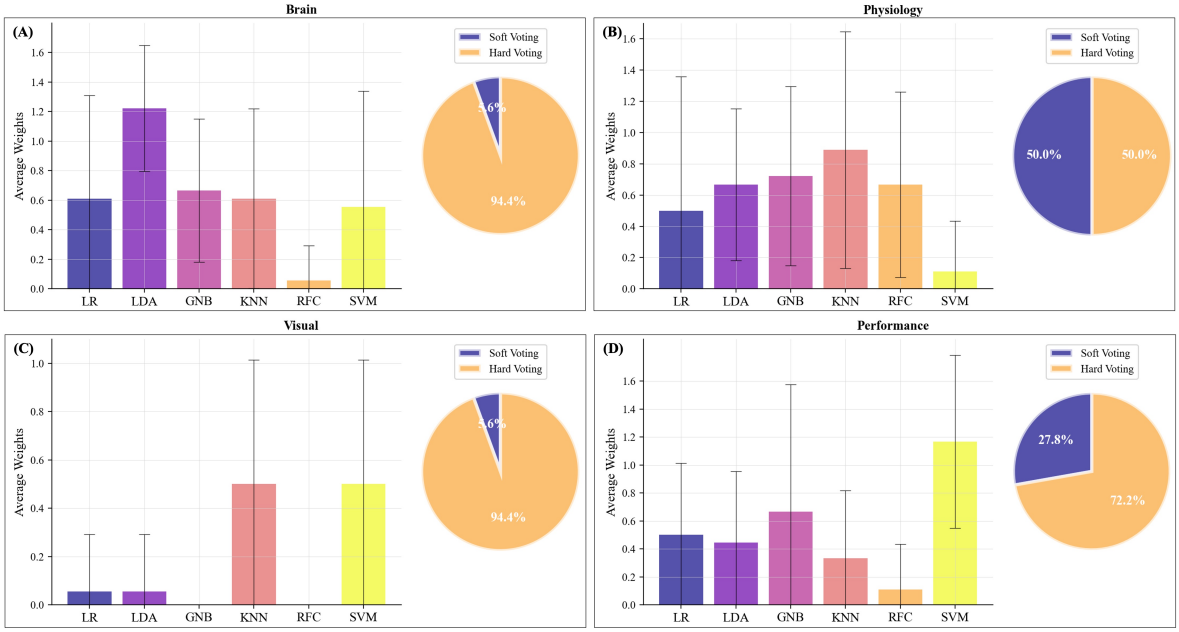

Figure 6. Weights and procedure of a unimodal voting classifier to predict subjectively perceived mental effort based on a median split; validation set: $N = 1$. Error bars represent the standard deviation.

Interestingly, for eight out of eighteen participants, we observed high prediction performances with F_1 scores ranging between 0.7 and 1.0. However, we also identified several subjects whose subjectively perceived mental effort was hard to predict based on the training data of the other subjects. See Table 1-3 in the Supplementary Material for a detailed comparison of the classifiers' performances in the different test subjects. Concluding, the results indicate that transfer learning and generalisation over subjects is much more challenging when using the neurophysiological compared with the performance-based features.

3.2. Unimodal Predictions – Brain Activity

The unimodal voting classifiers for brain activity mainly used hard voting (94.4%) and gave the highest weights to the LDA classifier. Classifiers revealed strong overfitting (Figure 5A) and neither a performance that was better than the single classifiers nor dummy classifier. We then compared the performance of the classifiers with respect to the percentage of correctly and falsely classified cases in a confusion matrix (Figure 7). Therefore, we used the best-performing classifier for each test subject and then summed over all test subjects. We compared the distribution of the true positives, true negatives, false positives, and false negatives in these classifiers with the respective distribution of the voting classifier. Here (Figure 7A), we see that both distributions indicate a high number of falsely identified "High Mental Effort" cases (False Positives), leading to a recall of 45.6% and precision of only 39.3% for the voting classifier and a recall of 57.5% and precision of 49.8% for single classifiers.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2023 doi:10.20944/preprints202306.1083.v1

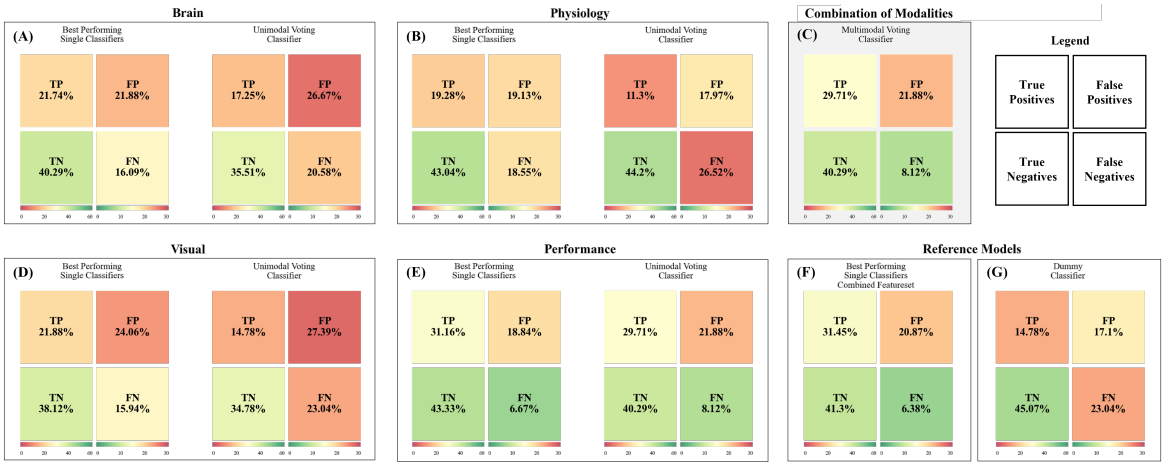

Figure 7. Prediction of the subjectively perceived mental effort (confusion matrix of test set) based on a median split; validation set: $N = 1$. Percentage of correctly and falsely classified perceived mental effort per model across all test subjects: TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives, with "Positives" representing "High Mental Effort" and "Negatives" representing "Low Mental Effort". For the "Best Performing Single Classifier" we selected the classifier (LDA, LR, SVM, KNN, RFC, or GNB) with the best F_1 score for each subject.

3.3. Unimodal Predictions – Physiological Measures

For classifying subjectively perceived mental effort based on physiological measures such as heart rate, respiration, and body temperature, soft voting was chosen in half of the test subjects (Figure 6B). The weighting of the classifiers varied considerably, with the KNN obtaining the highest average weights. The voting classifier (Figure 5B) showed strong overfitting, and its performance in the test subject was neither significantly better than any of the single classifiers nor dummy classifier. Regarding the percentage of correctly and falsely classified cases (Figure 7B), we see that the distributions for the best-performing single classifiers seem to be slightly better than the distributions of the voting classifier. The latter had difficulties in correctly identifying the conditions with low mental effort as can be seen in the high number of false negatives. When comparing the recall and precision of both approaches, we have a recall of only 29.9% for the voting classifier (precision: 38.6%) and an average recall of 51.0% for the best single classifiers (precision: 50.2%).

3.4. Unimodal Predictions – Ocular Measures

For subjectively perceived mental effort classification based on ocular measures such as pupil dilation and fixations, the split of soft vs. hard voting was 5.6% for soft voting and 94.4% for hard voting (Figure 6C). KNN and SVM were weighted highest. The F_1 score of the voting classifier (F_1 = .35, Figure 5D) did not show a significantly better classification performance than the dummy classifier (F_1 = .37). The percentage of correctly and falsely classified cases (Figure 7D) was similar to the brain models, with a recall of 39.1% for the voting classifier (precision: 35.1%) and an average recall of 57.9% for the best single classifiers (average precision: 47.6%).

3.5. Unimodal Predictions – Performance

At last, we predicted subjectively perceived mental effort based on performance (accuracy and speed). 27.8% of the test subjects had voting classifiers using soft voting, and 72.2% used hard voting (Figure 6D) with SVM being weighted highest. GNB, RFC, and SVM showed a significantly better performance than the dummy classifier. The performance of the combined voting classifier in the test subject was significantly better than a dummy classifier. The percentage of correctly and falsely classified cases (Figure 7E) reveals superior classification performance compared with the brain-, physiological- and ocular-based models. However, the voting classifier had still a high number of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 June 2023 doi:10.20944/preprints202306.1083.v1

falsely identified "High Mental Effort" cases (False Positives), leading to a recall of 78.5% and a precision of 57.6%. The best-performing single classifiers have an average recall of 82.4% and an average precision of 62.3%.

3.6. Unimodal Predictions based on the Upper Quartile Split

To identify informative measures for very high perceived mental effort potentially reflecting cognitive overload, we also performed predictions based on the subject-wise split at the upper quartile. Compared with the median-split-based results, we observed decreased classifiers' performance even below dummy classifier performance (Supplementary Figures 7). This might be explained by the fact that we reframed a binary prediction problem with evenly distributed classes into an outlier detection problem. Using the upper quartile split, we created imbalanced classes regarding the number of the respective samples, which made the reliable identification of the less well-represented class in the training set more difficult (reflected in the recall; Supplementary Figures 9).

3.7. Unimodal Predictions based on the Experimental Condition

We further fitted models to predict the experimentally induced task load instead of the subjectively perceived mental effort. The prediction of mental effort operationalised by the task load was substantially more successful than the prediction of subjectively perceived mental effort. All modalities, including brain activity and physiological activity, revealed at least one classifier that was able to predict the current task load above the chance level. The unimodal voting classifiers were all significantly better than a dummy classifier. Best unimodal voting classifications were obtained based on performance measures. Interestingly, other classification models were favoured in the unimodal voting, and the distribution between soft- and hard voting differed compared with the subjectively based approach, with soft voting being used more often (Supplementary Figure 17).

3.8. Multimodal Predictions based on the Median Split

In the final step, we combined the different modalities into a multimodal prediction. Figure 5C and Figure 7C shows the performance of the multimodal voting classifier, and Figure 8A the average allocated weights to the different modalities. To compare the rather complex feature set construction of the multimodal voting with a simpler approach, we also trained two exemplary classifiers (LR without feature selection and RFC with additional feature selection) on the whole feature set without a previous splitting into the different modalities (Figure 5F).

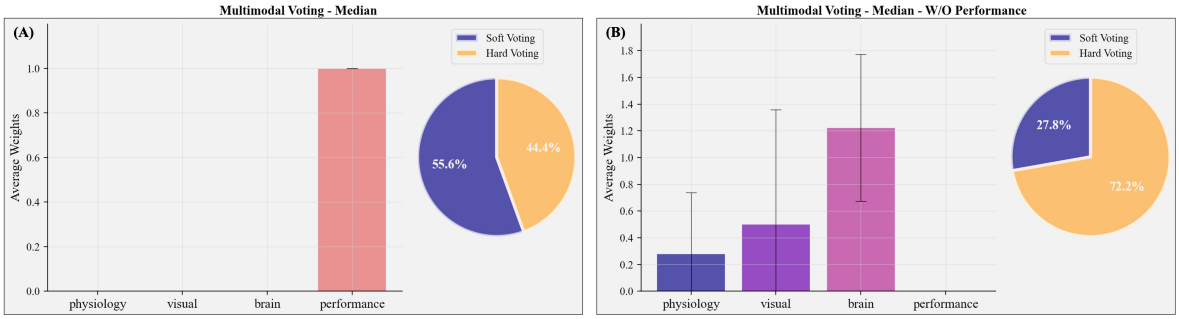

Figure 8. Weights and procedure of a multimodal voting classifier to predict subjectively perceived mental effort based on a median split; validation set: $N = 1$. (**B**) shows the allocation of weights when performance measures are not included in the multimodal classification. Error bars represent the standard deviation.

In most test subjects (55.6%), soft voting was selected to combine the predictions for the different modalities; 44.4% used hard voting. In line with the results outlined above, the multimodal classifier relied on performance to predict subjectively perceived mental effort, thereby turning it into a unimodal

 doi:10.20944/preprints202306.1083.v1

classifier. The voting classifier led to a significantly better classification than the dummy classifier (Figure 5C). The multimodal classifier exhibited an equivalent percentage of correctly and falsely classified cases (Figure 7C) compared with the performance-based classifier, demonstrating an average recall of 78.5% and an average precision of 57.6%. On average, it performed better than the classifiers trained with the combined whole feature set, which showed substantial overfitting.

In order to assess the performance of the multimodal classifier without incorporating performance-based information such as speed and accuracy, we constrained the classifier to utilize only (neuro-)physiological and visual measures. This approach is especially relevant for naturalistic applications where obtaining an accurate assessment of behavioural performance is challenging or impossible within the critical time window. For the multimodal prediction without performance, brain activity was weighted highest (Figure 8B). However, classifiers revealed strong overfitting during the training, and the average performance was decreased to chance level (average recall: 40.6% and average precision: 38.3%; Figure 5C).

3.9. Multimodal Predictions based on the Upper Quartile Split

With the upper quartile split, we observed a fundamentally different allocation of weights. High weights were assigned to brain and ocular activity (Figure 8B), while performance received only minimal weights. Hence, the exclusion of performance-based measures had minimal impact on the allocation of weights (Supplementary Figure 10B) and the overall performance of the multimodal classifiers remained largely unaffected (Supplementary Figure 7C). Among the eighteen test subjects, the multimodal classification demonstrated the highest performance in two cases (Supplementary Table 2). However, on average, the multimodal classification based on an upper quartile split did not demonstrate superiority over the unimodal classifiers. It further did not significantly outperform the dummy classifier or classifiers trained on a feature set of simply combined modalities without weight assignment (average recall: 21.9% and average precision: 18.5%; Supplementary Figure 7).

3.10. Multimodal Predictions based on the Experimental Condition

Similar to the multimodal voting classifier based on a subject-wise median split of perceived mental effort, classifiers predicted the experimentally induced task load solely using the performance measures. The average prediction performance was exceptionally high, significantly outperforming a dummy classifier, and comparable to the performance of the classifiers trained on the combined feature set (average recall: 99.7% and average precision: 91.3%; Supplementary Figure 16). When we only allowed (neuro-)physiological and visual measures as features, visual measures were weighted highest (Supplementary Figure 19B). In this case, the average performance of the multimodal classifiers was also significantly above the chance level, with an average recall of 82.7% and precision of 58.7%, indicating a successful identification of mental effort based on neurophysiological, physiological, and visual measures (Supplementary Figure 16C).

4. Discussion

The purpose of our study was to test the feasibility of multimodal voting in a machine learning classification for complex close-to-realistic scenarios. We used both - the subjectively experienced and experimentally induced mental effort - as ground truths for a cross-subject classification. Our approach represents a crucial investigation into the practical application of mental state decoding under real-world conditions. It serves as the foundation for enabling the adaptation of systems to users' current mental resources and efforts. By incorporating adaptive systems, individuals can enhance their performance by operating within an optimal level of demand, allowing them to perform at their best. In tasks involving high-security risks, it is crucial for system engineers to make every effort to prevent individuals from being overwhelmed or bored, as such states can increase the likelihood of errors. In the study presented here, we collected multimodal data from participants who performed a quasi-realistic experimental task involving varying levels of task complexity and salient distractions.

 doi:10.20944/preprints202306.1083.v1

In our analyses, we employed multimodal voting cross-subject classification and evaluated the model performance using a leave-one-out approach. Our results show which classifier models perform best for each modality. Furthermore, we observed that in certain modalities, the combination of ML models outperformed predictions made by individual ones. For each modality, we found a different set of classifiers that were better performing in the prediction and, thus, also considered more informative in the unimodal voting.

4.1. Using Subjectively Perceived Mental Effort as Ground Truth

When predicting subjectively perceived mental effort, LDA and LR performed best and were weighted highest in the classifications based on brain activity (Figure 5). Whereas, in physiological activity, the highest weights were assigned to KNN, RFC, and SVM. Regarding visual activity, the GNB and KNN revealed high classification performance among the test subjects. However, these models aiming to predict subjectively perceived mental effort based on brain activity, physiological activity, and visual measures were still strongly overfitted, and their performances in the test subjects were not significantly better than the dummy classifier. In performance-related measures, the GNB, RFC, and SVM performed significantly better than the dummy classifier when predicting subjective mental effort based on a median split. Using the upper quartile split for performance-related measures, the KNN and SVM showed the highest, but still, chance-level-like performances. Regarding the unimodal voting predictions of subjectively perceived mental effort, we see that a weighted combination of classifiers (LR, LDA, GNB, KNN, RFC, and SVM) was not superior to single classifiers neither when using the median nor the upper quartile split. When we combined the different modalities into a joined prediction of subjectively perceived mental effort, only the performance modality was considered. Hence, our multimodal classification might rather be considered a unimodal (performance-based) prediction. Removing the performance information from the multimodal voting classifier increases overfitting and drops the average classification performance. However, a more detailed investigation of the upper quartile split classification revealed that performance was less predictive in identifying cases of exceptionally high perceived mental effort and potential "cognitive overload" (Figure 7). In the upper quartile split classification higher multimodal voting weights were assigned to neurophysiological and visual measures compared with performance measures. This seems to imply that subjects were more heterogeneous in their performance under exceptionally high perceived mental effort, and classifiers rather exploited correlates from neurophysiological and visual measures than from performance to predict subjectively perceived mental effort. In summary, our findings indicate that when utilizing only unimodal voting classification, the best prediction of subjectively perceived mental effort was achieved through performance-based measures. Additionally, the inclusion of the performance-based classification model is essential in our multimodal voting classification approach to address potential overfitting in predicting mental effort (median-based split). These findings suggest that further research is necessary to investigate the dependence and variability of mental effort in cross-subject classification.

4.2. Using Experimentally Induced Mental Effort as Ground Truth

For the classification of the experimentally induced task load, all modalities were able to predict mental effort with high performances already on a single classifier level. GNB, KNN, and SVM performed above the chance level and were assigned the highest weights in the unimodal voting based on brain activity (Supplementary Figure 16). For the physiological activity, all classifiers - except the KNN - reached above-chance level performance. The highest average weight in the unimodal voting was assigned to the SVM. For visual and performance measures, we did not see substantial differences between the classification performance of the single models, with all performing above the chance level. A unimodal weighted combination of these classifiers was not superior to the single classifiers in any modality. Performance exhibited the highest predictive capability for task load (Supplementary Figure 16). As a result, the multimodal classifier transitioned again back into unimodal voting, as it relied solely on the performance modality. When excluding the performance-based features,

 doi:10.20944/preprints202306.1083.v1

the multimodal prediction based on neurophysiological, physiological, and ocular activity was still significantly above the chance level estimated by a dummy classifier. These findings suggest that it is feasible to differentiate between various mental effort states, represented by experimentally induced task load, by utilizing (neuro-)physiological and visual data obtained in a close-to-realistic environment through a cross-subject classification approach. However, it was not possible to replicate these results for subjectively perceived mental effort. The discrepancies observed between these two ground truth approaches could potentially be attributed to the retrospective nature of self-reports. Self-reports rely on an individual's perception, reasoning, and subjective introspection [78]. They are, therefore, vulnerable to various perceptual and response biases like social desirability [21,79]. These post-hoc evaluation processes might not be adequately reflected in and could be learned from (neuro-)physiological and visual measures during the task itself.

4.3. Generalisation across Subjects

For all classification approaches, we observed substantial variation in the performance of classifiers between the test subjects. Some individuals had F_1 scores above 0.8 (Supplementary Table 1). Other individuals demonstrated deviations in their neurophysiological reactions, diverging significantly from the patterns learned from the subjects included in the training set. These results are in line with the findings by Causse *et al.* [80]. The authors concluded that it is quite challenging to identify mental states based on haemodynamic activity across individuals because of the major structural and functional inter-individual differences. For instance, in the context of brain-computer interfaces, a phenomenon called BCI illiteracy describes the inability to modulate sensorimotor rhythms in order to control a BCI observed in approximately 20–30% of subjects [81]. Our findings underscore the importance of developing appropriate methods to address two key aspects. First, identifying subjects who may pose challenges in prediction due to their heterogeneity compared with the training set. Second, enabling transfer learning for these individuals by implementing techniques such as standardization and transformation of correlates into a unified feature space [82].

4.4. Limitations and Future Research

We acknowledge that certain aspects of this study can be further improved and serve as opportunities for future advancements. One area for improvement is the complexity of the measurement setup used in this study, which required a substantial amount of time for the preparation and calibration of the involved devices. It is important to consider the potential impact on participants' intrinsic engagement and explore ways to further streamline the process during soft- and hardware development. Furthermore, it is worth noting that our study sample was relatively homogeneous in terms of socio-demographic characteristics, consisting predominantly of young individuals with a high level of education. This homogeneity could potentially limit the generalizability of our results to more diverse populations. While it may seem intuitive to increase the sample size to address the issue of heterogeneity, there is a debate surrounding the relationship between sample size and its impact on classifier performance. With adding more and more samples, the dataset is supposedly at some point large enough to enable the classifier to find more generic and universal predictive patterns and achieve better performance again. Some argue that it is necessary to train ML models with large training datasets, including edge cases, to achieve good generalisability and attain good prediction accuracy on an individual-level Bzdok and Meyer-Lindenberg [83], Dwyer *et al.* [84]. Nevertheless, as emphasized by Cearns *et al.* [85], it is worth noting that machine learning classifiers demonstrate exceptional performance primarily in relatively small datasets. Consequently, the heterogeneity of a large dataset might present a significant challenge for learning. Thus, it may be more reasonable to train separate, specialized models for each homogeneous cluster, rather than attempting to construct a single model that explains the entire variance but yields less accurate predictions. Orrù *et al.* [86], for example, suggests the use of simple classifiers or ensemble learning methods instead of complex neural networks. Cearns *et al.* [85] highlight the importance of suitable cross-validation methods.

 doi:10.20944/preprints202306.1083.v1

Especially in the case of physiological datasets, one might also identify subjects that are very predictive for the patterns of a specific subgroup and remove subjects from the training set that show unusual patterns in neurophysiological reactions [84]. One interesting idea to address this problem is data augmentation [87,88]. This can be done by artificially generating new samples from existing samples to extend a dataset. For example, using Generative Adversarial Networks (GANs), one could simulate data to create more homogeneous and "prototypic" training datasets and increase the performance and stability of respective ML models [89]. Another suggested method to improve generalisability across subjects might be multiway canonical correlation analysis (MCCA). An approach that allows combining multiple data sets into a common representation and, thereby, achieves the denoising of data, and dimensionality reduction, based on shared components across subjects [82]. Advancements in these methods play a crucial role in enhancing the comparability and potential combinability of datasets, which is a shared objective within the research community [90].

To further increase classification performance, additional artefact analyses [91], or the implementation of inclusion criteria on a subject-, trial-, and channel-level could be explored in order to improve poor signal-to-noise ratios. Friedman *et al.* [92], who used an XGBoost classifier on EEG data, applied extensive and rigorous trial and subject selection criteria. For example, they did not include trials where participants failed to solve the task because they assumed that the mental effort shown by participants answering incorrectly did not reflect the true level of load (also [93]). Although this bears the risk of a major data loss, these rigid removal criteria might reflect an efficient solution to ensure that the measured neurophysiological signals truly reflect the cognitive processes of interest. Future research is necessary to a) define such exclusion and inclusion criteria depending on the investigated cognitive processes and b) develop standardised evaluation methods to decide which preprocessing step is beneficial and adequate.

A final limitation relates to the arrangement of the fNIRS optodes. Based on previous research (e.g., [12]), we decided to choose a montage solely covering the prefrontal cortex in order to reduce preparation time and facilitate transfer into close-to-realistic applications. However, we probably would have profited from a larger brain coverage that also covers parietal, temporal, and occipital brain areas [93]. Integrating these regions allows identifying features for the classification from larger functional networks that might play a crucial role in distinguishing mental states and cognitive control mechanisms [36,94]. Increased activity in the frontoparietal network is, for example, associated with task-related working memory (WM) processes (e.g., [95,96]), whereas increased connectivity between frontal and sensory areas are linked to the suppression of distractors [94].

4.5. Feature Selection and Data Fusion in Machine Learning

A crucial aim of this study was the selection and fusion of informative sources for cross-subject mental effort prediction. We integrated data from different modalities comprising brain activity as assessed with fNIRS, physiological activity (cardiac activity, respiration, and body temperature), ocular measures (pupil dilation and fixations), as well as behavioural measures of performance (accuracy and speed). However, this selection was naturally not exhaustive. Other measures, such as electroencephalography or electrodermal activity [97], could provide useful information about cognitive and physiological processes related to mental effort. In addition, one could also explore more behaviour-related measures such as speech [98] or gaze [99]. These measures might also provide the possibility to detect predictive patterns without significantly interfering with the actual task.

To combine the data streams obtained from the different measurement methods, we implemented data fusion on two levels: 1) the feature level and 2) the classification level. First, we aggregated our raw data, mainly time series, into informative features. We used standard statistical features like the mean, standard deviation, skewness, and kurtosis. Friedman *et al.* [92] explored more sophisticated features such as connectivity and complexity metrics, which have the potential to capture additional information about relationships within and between neuronal networks. Further investigations are required to assess the predictive quality of these aggregated features. Additionally, future research can explore

 doi:10.20944/preprints202306.1083.v1

the added value of feature selection and wrapping methods, which aim to reduce the complexity of the feature space without compromising the predictive information [100,101]. Such methods, e.g., sequential feature forward selection, might be a way to improve classifiers' performance by keeping only the most informative features. Another approach could be the use of continuous time-series data which provide insights into differences in the experience and processing of mentally demanding tasks separately for the different neurophysiological modalities. Hence, some researchers implemented deep learning methods like convolutional or recurrent neural networks to derive classifications based on multidimensional time-series data [45,102,103]. Nevertheless, these algorithms require that all data streams are complete (no missing data points) and have the same length and sampling frequency. These requirements are often difficult to fulfil in naturalistic settings with multimodal measurement methods using different measurement devices.

Once the feature space is defined, the research focus shifts towards developing strategies for selecting, merging, combining, and weighting multiple classifier models and modalities at the classification level. These strategies are still the subject of ongoing research and exploration. In this context, it is important to strike a balance between computational power, dataset size, and the benefits of finely tuned combinations of optimally stacked or voted classifiers. The exploration of early and late fusion approaches, as commonly employed in the field of robotics, could provide valuable insights. Early fusion involves the early combination of all data points and the fitting of classifiers to multidimensional data. On the other hand, late fusion involves a more fine-grained pipeline, where several classifiers are fitted to different proportions of the dataset and subsequently combined at a later stage. In this study, we implemented a late-fusion approach where we first combined different classifiers for each modality. In a subsequent step, we combined classifiers to create a unified prediction. Exploring early and late fusion strategies is especially important when one wants to account for temporal dynamics in the different measures or the realisation of real-time mental state monitoring. The review of Debie *et al.* [27] provides a comprehensive overview of the different fusion stages when identifying mental effort based on neurophysiological measures.

5. Practical Implications and Conclusion

Our proposed multimodal voting classification approach contributes to the ecologically valid distinction and identification of different states of mental effort. It paves the way toward generalised state monitoring across individuals in realistic applications. Interestingly, the choice of ground truth had a fundamental influence on the classification performance. The prediction of subjectively perceived mental effort operationalized through self-reports, is most effectively achieved by incorporating performance-based measures. On the other hand, the experimentally induced task load can be accurately predicted not only from performance-based measures but also by incorporating neurophysiological and visual measures. Our findings provide valuable guidance for researchers and practitioners in selecting appropriate methods based on their specific research questions or application scenarios, taking into account limited resources or environmental constraints. The capacity to predict subjectively perceived and experimentally induced mental effort on an individual level makes this architecture an integral part of future research and development of user-centred applications such as adaptive assistance systems.

Supplementary Materials: The supplementary material can be downloaded at https://osf.io/9dbcj/files/osfstorage/6450b885d805c504d75522ef.

Author Contributions: Conceptualization, Katharina Lingelbach; Data curation, Sabrina Gado and Katharina Lingelbach; Formal analysis, Sabrina Gado; Funding acquisition, Maria Wirzberger; Methodology, Katharina Lingelbach; Project administration, Katharina Lingelbach; Supervision, Katharina Lingelbach and Mathias Vukelić; Visualization, Sabrina Gado; Writing – original draft, Sabrina Gado; Writing – review & editing, Sabrina Gado, Katharina Lingelbach, Maria Wirzberger and Mathias Vukelić.

Funding: The reported research was supported by the Federal Ministry of Science, Research, and the Arts Baden-Württemberg and the University of Stuttgart as part of the Research Seed Capital funding scheme as

well as by a grant from the Ministry of Economic Affairs, Labour and Tourism Baden-Wuerttemberg (Project »KI-Fortschrittszentrum Lernende Systeme und Kognitive Robotik«).

Institutional Review Board Statement: This study was approved by the ethics committee of the Medical Faculty of the University of Tuebingen, Germany (ID: 827/2020BO1).

Informed Consent Statement: Participants were informed that their participation was voluntary and that they could withdraw at any time during the experiment. They signed an informed consent according to the recommendations of the Declaration of Helsinki.

Data Availability Statement: The datasets analysed for this study as well as the code can be found in a publicly accessible OSF repository: https://osf.io/9dbcj/.

Acknowledgments: We would like to thank Ron Becker, Alina Schmitz-Hübsch, Michael Bui, and Sophie Felicitas Böhm for their contribution to the experimental environment, technical set-up, data collection and data preparation.

Conflicts of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

fNIRS	functional Near-Infrared Spectroscopy
ECG	Electrocardiography
HbO	Oxy-Haemoglobin
HbR	Deoxy-Haemoglobin
PFC	Prefrontal Cortex
WCT	Warship Commander Task
SD	Standard Deviation
CI	Confidence Interval
ML	Machine Learning
LR	Logistic Regression
LDA	Linear Discriminant Analysis
GNB	Gaussian Naïve Bayes Classifier
KNN	K-Nearest Neighbor Classifier
RFC	Random Forest Classifier
SVM	Support Vector Machine Classifier

References

1. Hart, S.G.; Staveland, L.E. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in Psychology*; Hancock, P.A.; Meshkati, N., Eds.; North-Holland, 1988; Vol. 52, pp. 139–183. doi:10.1016/S0166-4115(08)62386-9.
2. Young, M.S.; Brookhuis, K.A.; Wickens, C.D.; Hancock, P.A. State of science: Mental workload in ergonomics. *Ergonomics* **2015**, *58*, 1–17. doi:10.1080/00140139.2014.956151.
3. Paas, F.; Tuovinen, J.E.; Tabbers, H.; Van Gerven, P.W.M. Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist* **2003**, *38*, 63–71. Publisher: Routledge, doi:10.1207/S15326985EP3801_8.
4. Chen, F.; Zhou, J.; Wang, Y.; Yu, K.; Arshad, S.Z.; Khawaji, A.; Conway, D. *Robust multimodal cognitive load measurement*; Human–Computer Interaction Series, Springer International Publishing: Cham, 2016. doi:10.1007/978-3-319-31700-7.
5. Zheng, R.Z. *Cognitive load measurement and application: A theoretical framework for meaningful research and practice*; Routledge: New York, NY, US, 2017. Pages: xiii, 278.
6. von Lühmann, A. Multimodal instrumentation and methods for neurotechnology out of the lab. Doctoral Dissertation, Technische Universität Berlin, Berlin, 2018. Publication Title: Fakultät IV - Elektrotechnik und Informatik, doi:10.14279/depositonce-7445.
7. Charles, R.L.; Nixon, J. Measuring mental workload using physiological measures: A systematic review. *Applied Ergonomics* **2019**, *74*, 221–232. doi:10.1016/j. apergo.2018.08.028.

8. Curtin, A.; Ayaz, H. The age of neuroergonomics: Towards ubiquitous and continuous measurement of brain function with fNIRS. *Japanese Psychological Research* **2018**, *60*, 374–386. Publisher: Wiley-Blackwell Publishing Ltd. Blackwell Publishing, doi:10.1111/jpr.12227.
9. Benerradi, J.; Maior, H.A.; Marinescu, A.; Clos, J.; Wilson, M.L. Mental workload using fNIRS data from HCI tasks ground truth: Performance, evaluation, or condition. Proceedings of the Halfway to the Future Symposium; Association for Computing Machinery: Nottingham, United Kingdom, 2019. Type: 10.1145/3363384.3363392, doi:10.1145/3363384.3363392.
10. Midha, S.; Maior, H.A.; Wilson, M.L.; Sharples, S. Measuring mental workload variations in office work tasks using fNIRS. *International Journal of Human-Computer Studies* **2021**, *147*, 102580. doi:10.1016/j.ijhcs.2020.102580.
11. Izzetoglu, K.; Bunce, S.; Izzetoglu, M.; Onaral, B.; Pourrezaei, K. fNIR spectroscopy as a measure of cognitive task load. Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2003, Vol. 4, pp. 3431–3434 Vol.4. Journal Abbreviation: Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE Cat. No.03CH37439), doi:10.1109/IEMBS.2003.1280883.
12. Ayaz, H.; Shewokis, P.A.; Bunce, S.; Izzetoglu, K.; Willems, B.; Onaral, B. Optical brain monitoring for operator training and mental workload assessment. *NeuroImage* **2012**, *59*, 36–47. doi:10.1016/j.neuroimage.2011.06.023.
13. Herff, C.; Heger, D.; Fortmann, O.; Hennrich, J.; Putze, F.; Schultz, T. Mental workload during n-back task - Quantified in the prefrontal cortex using fNIRS. *Frontiers in Human Neuroscience*, *7*, 935. doi:10.3389/fnhum.2013.00935.
14. Miller, E.K.; Freedman, D.J.; Wallis, J.D. The prefrontal cortex: Categories, concepts and cognition. *Philosophical Transactions of the Royal Society of London* **2002**, *357*, 1123–1136. doi:10.1098/rstb.2002.1099.
15. Dehais, F.; Lafont, A.; Roy, R.; Fairclough, S. A neuroergonomics approach to mental workload, engagement and human performance. *Frontiers in Neuroscience* **2020**, *14*, 268. doi:10.3389/fnins.2020.00268.
16. Babiloni, F. Mental workload monitoring: New perspectives from neuroscience. Human Mental Workload: Models and Applications; Longo, L.; Leva, M.C., Eds.; Springer International Publishing: Cham, 2019; Vol. 1107, *Communications in Computer and Information Science*, pp. 3–19. doi:10.1007/978-3-030-32423-0_1.
17. Matthews, R.; McDonald, N.J.; Trejo, L.J. Psycho-physiological sensor techniques: An overview. In *Foundations of Augmented Cognition*; Schmorrow, D.D., Ed.; CRC Press, 2005; Vol. 11, pp. 263–272. doi:10.1201/9781482289701.
18. Wierwille, W.W. Physiological measures of aircrew mental workload. *Human Factors* **1979**, *21*, 575–593. doi:10.1177/001872087902100504.
19. Kramer, A.F. Physiological metrics of mental workload: A review of recent progress. In *Multiple-task performance*; Damos, D.L., Ed.; CRC Press: London, 1991; pp. 279–328. doi:10.1201/9781003069447.
20. Backs, R.W. Application of psychophysiological models to mental workload. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* **2000**, *44*, 464–467. doi:10.1177/154193120004402123.
21. Dirican, A.C.; Göktürk, M. Psychophysiological measures of human cognitive states applied in human computer interaction. *Procedia Computer Science* **2011**, *3*, 1361–1367. doi:10.1016/j. procs.2011.01.016.
22. Dan, A.; Reiner, M. Real time EEG based measurements of cognitive load indicates mental states during learning. *Journal of Educational Data Mining* **2017**, *9*, 31–44. doi:10.5281/zenodo.3554719.
23. Tao, D.; Tan, H.; Wang, H.; Zhang, X.; Qu, X.; Zhang, T. A systematic review of physiological measures of mental workload. *International Journal of Environmental Research and Public Health* **2019**, *16*, 2716. doi:10.3390/ijerph16152716.
24. Romine, W.L.; Schroeder, N.L.; Graft, J.; Yang, F.; Sadeghi, R.; Zabihimayvan, M.; Kadariya, D.; Banerjee, T. 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* **2020**, *20*. doi:10.3390/s20174833.
25. Uludağ, K.; Roebroeck, A. General overview on the merits of multimodal neuroimaging data fusion. *NeuroImage* **2014**, *102*, 3–10. doi:10.1016/j.neuroimage.2014.05.018.
26. Zhang, Y.D.; Dong, Z.; Wang, S.H.; Yu, X.; Yao, X.; Zhou, Q.; Hu, H.; Li, M.; Jiménez-Mesa, C.; Ramirez, J.; Martinez, F.J.; Gorriz, J.M. Advances in multimodal data fusion in neuroimaging: Overview, challenges, and novel orientation. *Information Fusion* **2020**, *64*, 149–187. doi:10.1016/j.inffus.2020.07.006.

27. Debie, E.; Rojas, R.F.; Fidock, J.; Barlow, M.; Kasmarik, K.; Anavatti, S.; Garratt, M.; Abbass, H.A. Multimodal fusion for objective assessment of cognitive workload: A review. *IEEE Transactions on Cybernetics* **2021**, *51*, 1542–1555. doi:10.1109/TCYB.2019.2939399.
28. Klimesch, W. Evoked alpha and early access to the knowledge system: The P1 inhibition timing hypothesis. *Brain Research* **2011**, *1408*, 52–71. doi:10.1016/j.brainres.2011.06.003.
29. Wirzberger, M.; Herms, R.; Esmaeili Bijarsari, S.; Eibl, M.; Rey, G.D. Schema-related cognitive load influences performance, speech, and physiology in a dual-task setting: A continuous multi-measure approach. *Cognitive Research: Principles and Implications* **2018**, *3*, 46. doi:10.1186/s41235-018-0138-z.
30. Lemm, S.; Blankertz, B.; Dickhaus, T.; Müller, K.R. Introduction to machine learning for brain imaging. *Multivariate Decoding and Brain Reading* **2011**, *56*, 387–399. doi:10.1016/j.neuroimage.2010.11.004.
31. Vu, M.A.T.; Adalı, T.; Ba, D.; Buzsáki, G.; Carlson, D.; Heller, K.; Liston, C.; Rudin, C.; Sohal, V.S.; Widge, A.S.; Mayberg, H.S.; Sapiro, G.; Dzirasa, K. A shared vision for machine learning in neuroscience. *The Journal of Neuroscience* **2018**, *38*, 1601. doi:10.1523/JNEUROSCI.0508-17.2018.
32. Herms, R.; Wirzberger, M.; Eibl, M.; Rey, G.D. CoLoSS: Cognitive load corpus with speech and performance data from a symbol-digit dual-task. Proceedings of the 11th International Conference on Language Resources and Evaluation; European Language Resources Association: Miyazaki, Japan, 2018.
33. Ladouce, S.; Donaldson, D.I.; Dudchenko, P.A.; Ietswaart, M. Understanding minds in real-world environments: Toward a mobile cognition approach. *Frontiers in Human Neuroscience* **2017**, *10*, 694. doi:10.3389/fnhum.2016.00694.
34. Lavie, N. Attention, distraction, and cognitive control under load. *Current Directions in Psychological Science* **2010**, *19*, 143–148. Publisher: SAGE Publications Inc, doi:10.1177/0963721410370295.
35. Baddeley, A.D.; Hitch, G. Working Memory. In *Psychology of Learning and Motivation*; Bower, G.H., Ed.; Academic Press, 1974; Vol. 8, pp. 47–89. doi:10.1016/S0079-7421(08)60452-1.
36. Soerqvist, P.; Dahlstroem, O.; Karlsson, T.; Rönnberg, J. Concentration: The neural underpinnings of how cognitive load shields against distraction. *Frontiers in Human Neuroscience* **2016**, *10*, 221. doi:10.3389/fnhum.2016.00221.
37. Anikin, A. The link between auditory salience and emotion intensity. *Cognition and Emotion* **2020**, *34*, 1246–1259. Publisher: Routledge _eprint: https://doi.org/10.1080/02699931.2020.1736992, doi:10.1080/02699931.2020.1736992.
38. Dolcos, F.; Iordan, A.D.; Dolcos, S. Neural correlates of emotion–cognition interactions: A review of evidence from brain imaging investigations. *Journal of Cognitive Psychology* **2011**, *23*, 669–694.
39. D'Andrea-Penna, G.M.; Frank, S.M.; Heatherton, T.F.; Tse, P.U. Distracting tracking: Interactions between negative emotion and attentional load in multiple-object tracking. *Emotion* **2017**, *17*, 900–904. Place: US Publisher: American Psychological Association, doi:10.1037/emo0000329.
40. Schweizer, S.; Satpute, A.B.; Atzil, S.; Field, A.P.; Hitchcock, C.; Black, M.; Barrett, L.F.; Dalgleish, T. The impact of affective information on working memory: A pair of meta-analytic reviews of behavioral and neuroimaging evidence. *Psychological Bulletin* **2019**, *145*, 566–609. Place: US Publisher: American Psychological Association, doi:10.1037/bul0000193.
41. Banbury, S.; Berry, D.C. Disruption of office-related tasks by speech and office noise. *British Journal of Psychology* **1998**, *89*, 499–517. Publisher: John Wiley & Sons, Ltd, doi:10.1111/j.2044-8295.1998.tb02699.x.
42. Liebl, A.; Haller, J.; Jödicke, B.; Baumgartner, H.; Schlittmeier, S.; Hellbrück, J. Combined effects of acoustic and visual distraction on cognitive performance and well-being. *Applied Ergonomics* **2012**, *43*, 424–434. doi:10.1016/j. apergo.2011.06.017.
43. Vuilleumier, P.; Schwartz, S. Emotional facial expressions capture attention. *Neurology* **2001**, *56*, 153–158. Publisher: Wolters Kluwer Health, Inc. on behalf of the American Academy of Neurology Section: Articles, doi:10.1212/WNL.56.2.153.
44. Waytowich, N.R.; Lawhern, V.J.; Bohannon, A.W.; Ball, K.R.; Lance, B.J. Spectral transfer learning using information geometry for a user-independent brain-computer interface. *Frontiers in Neuroscience* **2016**, *10*, 430. doi:10.3389/fnins.2016.00430.
45. Lawhern, V.J.; Solon, A.J.; Waytowich, N.R.; Gordon, S.M.; Hung, C.P.; Lance, B.J. EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering* **2018**, *15*, 056013. Publisher: IOP Publishing, doi:10.1088/1741-2552/ aace8c.

46. Lyu, B.; Pham, T.; Blaney, G.; Haga, Z.; Sassaroli, A.; Fantini, S.; Aeron, S. Domain adaptation for robust workload level alignment between sessions and subjects using fNIRS. *Journal of Biomedical Optics* **2021**, *26*, 1–21. doi:10.1117/1.JBO.26.2.022908.
47. Lotte, F.; Bougrain, L.; Cichocki, A.; Clerc, M.; Congedo, M.; Rakotomamonjy, A.; Yger, F. A review of classification algorithms for EEG-based brain–computer interfaces: A 10 year update. *Journal of Neural Engineering* **2018**, *15*, 031005. Publisher: IOP Publishing, doi:10.1088/1741-2552/ aab2f2.
48. Liu, Y.; Lan, Z.; Cui, J.; Sourina, O.; Müller-Wittig, W. EEG-based cross-subject mental fatigue recognition. Proceedings of the International Conference on Cyberworlds 2019, 2019, pp. 247–252. Journal Abbreviation: 2019 International Conference on Cyberworlds (CW), doi:10.1109/CW.2019.00048.
49. Becker, R.; Stasch, S.M.; Schmitz-Hübsch, A.; Fuchs, S. Quantitative scoring system to assess performance in experimental environments. Proceedings of the 14th International Conference on Advances in Computer-Human Interactions; ThinkMind: Nice, France, 2021; pp. 91–96.
50. Burkhardt, F.; Paeschke, A.; Rolfes, M.; Sendlmeier, W.; Weiss, B. A database of German emotional speech. Proceedings of the 9th European Conference on Speech Communication and Technology; , 2005; Vol. 5, p. 1520. Journal Abbreviation: 9th European Conference on Speech Communication and Technology Publication Title: 9th European Conference on Speech Communication and Technology DOI:, doi:10.21437/Interspeech.2005-446.
51. Group, P.S..E. Warship Commander 4.4, 2003.
52. St John, M.; Kobus, D.A.; Morrison, J.G. DARPA augmented cognition technical integration experiment (TIE). Technical Report ADA420147, Pacific Science and Engineering Group, San Diego, CA, USA, 2003.
53. Toet, A.; Kaneko, D.; Ushiama, S.; Hoving, S.; de Kruijf, I.; Brouwer, A.M.; Kallen, V.; van Erp, J.B.F. EmojiGrid: A 2D pictorial scale for the assessment of food elicited emotions. *Frontiers in Psychology* **2018**, *9*, 2396. doi:10.3389/fpsyg.2018.02396.
54. Cardoso, B.; Romão, T.; Correia, N. CAAT: A discrete approach to emotion assessment. Extended Abstracts on Human Factors in Computing Systems; Association for Computing Machinery: Paris, France, 2013; pp. 1047–1052. Type: 10.1145/2468356.2468543, doi:10.1145/2468356.2468543.
55. Rammstedt, B.; John, O.P. Kurzversion des Big Five Inventory (BFI-K):. *Diagnostica* **2005**, *51*, 195–206. Publisher: Hogrefe Verlag, doi:10.1026/0012-1924.51.4.195.
56. Laux, L.; Glanzmann, P.; Schaffner, P.; Spielberger, C.D. *Das State-Trait-Angstinventar*; Beltz: Weinheim, 1981.
57. Bankstahl, U.; Görtelmeyer, R. *APSA: Attention and Performance Self-Assessment - deutsche Fassung*; Elektronisches Testarchiv, ZPID (Leibniz Institute for Psychology Information)–Testarchiv: Trier, 2013. [Fragebogen].
58. Hartmann, A.S.; Rief, W.; Hilbert, A. Psychometric properties of the German version of the Barratt Impulsiveness Scale, Version 11 (BIS–11) for adolescents. *Perceptual and Motor Skills* **2011**, *112*, 353–368. Publisher: SAGE Publications Inc, doi:10.2466/08.09.10.PMS.112.2.353-368.
59. Scheunemann, J.; Unni, A.; Ihme, K.; Jipp, M.; Rieger, J.W. Demonstrating brain-level interactions between visuospatial attentional demands and working memory load while driving using functional near-infrared spectroscopy. *Frontiers in Human Neuroscience* **2019**, *12*, 542. doi:10.3389/fnhum.2018.00542.
60. Zimeo Morais, G.A.; Balardin, J.B.; Sato, J.R. fNIRS Optodes' Location Decider (fOLD): A toolbox for probe arrangement guided by brain regions-of-interest. *Scientific Reports* **2018**, *8*, 3341. doi:10.1038/s41598-018-21716-z.
61. Dink, J.W.; Ferguson, B. eyetrackingR: An R library for eye-tracking data analysis, 2015.
62. Forbes, S. PupillometryR: An R package for preparing and analysing pupillometry data. *Journal of Open Source Software* **2020**, *5*, 2285. doi:10.21105/joss.02285.
63. Jackson, I.; Sirois, S. Infant cognition: Going full factorial with pupil dilation. *Developmental Ccience* **2009**, *12*, 670–679. Place: England, doi:10.1111/j.1467-7687.2008.00805.x.
64. von der Malsburg, T. saccades: Detection of fixations in eye-tracking data, 2015.
65. Gramfort, A.; Luessi, M.; Larson, E.; Engemann, D.A.; Strohmeier, D.; Brodbeck, C.; Parkkonen, L.; Hämäläinen, M.S. MNE Software for Processing MEG and EEG Data. *NeuroImage* **2014**, *86*, 446–460. doi:10.1016/j. neuroimage.2013.10.027.

66. Luke, R.; Larson, E.D.; Shader, M.J.; Innes-Brown, H.; Van Yper, L.; Lee, A.K.C.; Sowman, P.F.; McAlpine, D. Analysis methods for measuring passive auditory fNIRS responses generated by a block-design paradigm. *Neurophotonics* **2021**, *8*, 1–18. doi:10.1117/1.NPh.8.2.025008.
67. Yücel, M.A.; Lühmann, A.v.; Scholkmann, F.; Gervain, J.; Dan, I.; Ayaz, H.; Boas, D.; Cooper, R.J.; Culver, J.; Elwell, C.E.; Eggebrecht, A.; Franceschini, M.A.; Grova, C.; Homae, F.; Lesage, F.; Obrig, H.; Tachtsidis, I.; Tak, S.; Tong, Y.; Torricelli, A.; Wabnitz, H.; Wolf, M. Best practices for fNIRS publications. *Neurophotonics* **2021**, *8*, 1–34. doi:10.1117/1. NPh.8.1.012101.
68. Pollonini, L.; Olds, C.; Abaya, H.; Bortfeld, H.; Beauchamp, M.S.; Oghalai, J.S. Auditory cortex activation to natural speech and simulated cochlear implant speech measured with functional near-infrared spectroscopy. *Hearing Research* **2014**, *309*, 84–93. doi:10.1016/j. heares.2013.11.007.
69. Fishburn, F.A.; Ludlum, R.S.; Vaidya, C.J.; Medvedev, A.V. Temporal Derivative Distribution Repair (TDDR): A motion correction method for fNIRS. *NeuroImage* **2019**, *184*, 171–179. doi:10.1016/j. neuroimage.2018.09.025.
70. Saager, R.B.; Berger, A.J. Direct characterization and removal of interfering absorption trends in two-layer turbid media. *Journal of the Optical Society of America A* **2005**, *22*, 1874–1882. Publisher: OSA, doi:10.1364/JOSAA.22.001874.
71. Beer, A. Bestimmung der Absorption des rothen Lichts in farbigen Flüssigkeiten. *Annalen der Physik und Chemie* **1852**, *86*, 78–88.
72. Schiratti, J.B.; Le Douget, J.E.; Le Van Quyen, M.; Essid, S.; Gramfort, A. An ensemble learning approach to detect epileptic seizures from long intracranial EEG recordings. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing 2018, 2018, pp. 856–860. Journal Abbreviation: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), doi:10.1109/ICASSP.2018.8461489.
73. Keles, H.O.; Cengiz, C.; Demiral, I.; Ozmen, M.M.; Omurtag, A. High density optical neuroimaging predicts surgeons's subjective experience and skill levels. *PLOS ONE* **2021**, *16*, e0247117. Publisher: Public Library of Science, doi:10.1371/journal.pone.0247117.
74. Minkley, N.; Xu, K.M.; Krell, M. Analyzing relationships between causal and assessment factors of cognitive load: Associations between objective and subjective measures of cognitive load, stress, interest, and self-concept. *Frontiers in Education* **2021**, *6*.
75. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.
76. Raschka, S. MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack. *Journal of Open Source Software* **2018**, *3*, 638. doi:10.21105/joss.00638.
77. Cumming, G.; Finch, S. Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist* **2005**, *60*, 170–180. Place: US Publisher: American Psychological Association, doi:10.1037/0003-066X.60.2.170.
78. Ranchet, M.; Morgan, J.C.; Akinwuntan, A.E.; Devos, H. Cognitive workload across the spectrum of cognitive impairments: A systematic review of physiological measures. *Neuroscience & Biobehavioral Reviews* **2017**, *80*, 516–537. doi:10.1016/j. neubiorev.2017.07.001.
79. Matthews, G.; De Winter, J.; Hancock, P.A. What do subjective workload scales really measure? Operational and representational solutions to divergence of workload measures. *Theoretical Issues in Ergonomics Science* **2020**, *21*, 369–396. doi:10.1080/1463922X.2018.1547459.
80. Causse, M.; Chua, Z.; Peysakhovich, V.; Del Campo, N.; Matton, N. Mental workload and neural efficiency quantified in the prefrontal cortex using fNIRS. *Scientific Reports* **2017**, *7*, 5222. doi:10.1038/s41598-017-05378-x.
81. Allison, B.Z.; Neuper, C. Could anyone use a BCI? In *Brain-Computer Interfaces: Applying our Minds to Human-Computer Interaction*; Tan, D.S.; Nijholt, A., Eds.; Springer: London, 2010; pp. 35–54. doi:10.1007/978-1-84996-272-8_3.
82. de Cheveigné, A.; Di Liberto, G.M.; Arzounian, D.; Wong, D.D.; Hjortkjær, J.; Fuglsang, S.; Parra, L.C. Multiway canonical correlation analysis of brain data. *NeuroImage* **2019**, *186*, 728–740. doi:10.1016/j. neuroimage.2018.11.026.

83. Bzdok, D.; Meyer-Lindenberg, A. Machine learning for precision psychiatry: Opportunities and challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* **2018**, *3*, 223–230. doi:10.1016/j.bpsc.2017.11.007.
84. Dwyer, D.B.; Falkai, P.; Koutsouleris, N. Machine learning approaches for clinical psychology and psychiatry. *Annual Review of Clinical Psychology* **2018**, *14*, 91–118. doi:10.1146/annurev-clinpsy-032816-045037.
85. Cearns, M.; Hahn, T.; Baune, B.T. Recommendations and future directions for supervised machine learning in psychiatry. *Translational Psychiatry* **2019**, *9*, 271. doi:10.1038/s41398-019-0607-2.
86. Orrù, G.; Monaro, M.; Conversano, C.; Gemignani, A.; Sartori, G. Machine learning in psychometrics and psychological research. *Frontiers in Psychology* **2020**, *10*, 2970. doi:10.3389/fpsyg.2019.02970.
87. Lashgari, E.; Liang, D.; Maoz, U. Data augmentation for deep-learning-based electroencephalography. *Journal of Neuroscience Methods* **2020**, *346*, 108885. doi:10.1016/j.jneumeth.2020.108885.
88. Bird, J.J.; Pritchard, M.; Fratini, A.; Ekárt, A.; Faria, D.R. Synthetic biological signals machine-generated by GPT-2 improve the classification of EEG and EMG through data augmentation. *IEEE Robotics and Automation Letters* **2021**, *6*, 3498–3504. doi:10.1109/LRA.2021.3056355.
89. Zanini, R.A.; Colombini, E.L. Parkinson's disease EMG data augmentation and simulation with DCGANs and Style Transfer. *Sensors* **2020**, *20*, 2605. doi:10.3390/s20092605.
90. Abrams, M.B.; Bjaalie, J.G.; Das, S.; Egan, G.F.; Ghosh, S.S.; Goscinski, W.J.; Grethe, J.S.; Kotaleski, J.H.; Ho, E.T.W.; Kennedy, D.N.; Lanyon, L.J.; Leergaard, T.B.; Mayberg, H.S.; Milanesi, L.; Mouček, R.; Poline, J.B.; Roy, P.K.; Strother, S.C.; Tang, T.B.; Tiesinga, P.; Wachtler, T.; Wójcik, D.K.; Martone, M.E. A standards organization for open and FAIR neuroscience: The international neuroinformatics coordinating facility. *Neuroinformatics* **2021**. doi:10.1007/s12021-020-09509-0.
91. von Lühmann, A.; Boukouvalas, Z.; Müller, K.R.; Adalı, T. A new blind source separation framework for signal analysis and artifact rejection in functional near-infrared spectroscopy. *NeuroImage* **2019**, *200*, 72–88. doi:10.1016/j. neuroimage.2019.06.021.
92. Friedman, N.; Fekete, T.; Gal, K.; Shriki, O. EEG-based prediction of cognitive load in intelligence tests. *Frontiers in Human Neuroscience* **2019**, *13*, 191. doi:10.3389/fnhum.2019.00191.
93. Unni, A.; Ihme, K.; Jipp, M.; Rieger, J.W. Assessing the driver's current level of working memory load with high density functional near-infrared spectroscopy: A realistic driving simulator study. *Frontiers in Human Neuroscience* **2017**, *11*, 167. doi:10.3389/fnhum.2017.00167.
94. García-Pacios, J.; Garcés, P.; del Río, D.; Maestú, F. Tracking the effect of emotional distraction in working memory brain networks: Evidence from an MEG study. *Psychophysiology* **2017**, *54*, 1726–1740. Publisher: John Wiley & Sons, Ltd, doi:10.1111/psyp.12912.
95. Curtis, C. Prefrontal and parietal contributions to spatial working memory. *Neuroscience* **2006**, *139*, 173–180. doi:10.1016/j.neuroscience.2005.04.070.
96. Martínez-Vázquez, P.; Gail, A. Directed interaction between monkey premotor and posterior parietal cortex during motor-goal retrieval from working memory. *Cerebral Cortex* **2018**, *28*, 1866–1881. doi:10.1093/cercor/bhy035.
97. Vanneste, P.; Raes, A.; Morton, J.; Bombeke, K.; Van Acker, B.B.; Larmuseau, C.; Depaepe, F.; Van den Noortgate, W. Towards measuring cognitive load through multimodal physiological data. *Cognition, Technology & Work* **2021**, *23*, 567–585. doi:10.1007/s10111-020-00641-0.
98. Yap, T.F.; Epps, J.; Ambikairajah, E.; Choi, E.H. Voice source under cognitive load: Effects and classification. *Speech Communication* **2015**, *72*, 74–95. doi:10.1016/j. specom.2015.05.007.
99. Marquart, G.; Cabrall, C.; de Winter, J. Review of eye-related measures of drivers' mental workload. *Procedia Manufacturing* **2015**, *3*, 2854–2861. doi:10.1016/j. promfg.2015.07.783.
100. Gottemukkula, V.; Derakhshani, R. Classification-guided feature selection for NIRS-based BCI. Proceedings of the 5th International IEEE/EMBS Conference on Neural Engineering 2011, 2011, pp. 72–75. Journal Abbreviation: 2011 5th International IEEE/EMBS Conference on Neural Engineering, doi:10.1109/NER.2011.5910491.
101. Aydin, E.A. Subject-Specific feature selection for near infrared spectroscopy based brain-computer interfaces. *Computer Methods and Programs in Biomedicine* **2020**, *195*, 105535. doi:10.1016/j.cmpb.2020.105535.

102. Chakraborty, S.; Aich, S.; Joo, M.i.; Sain, M.; Kim, H.C. A multichannel convolutional neural network architecture for the detection of the state of mind using physiological signals from wearable devices. *Journal of Healthcare Engineering* **2019**, *2019*, 5397814. Publisher: Hindawi, doi:10.1155/2019/5397814.
103. Asgher, U.; Khalil, K.; Khan, M.J.; Ahmad, R.; Butt, S.I.; Ayaz, Y.; Naseer, N.; Nazir, S. Enhanced accuracy for multiclass mental workload detection using long short-term memory for brain–computer interface. *Frontiers in Neuroscience* **2020**, *14*, 584. doi:10.3389/fnins.2020.00584.