# Classifying Mental Effort in a Quasi-Realistic Scenario Based on Multimodal Data Fusion

**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 Wuerzburg, Wuerzburg, Germany

[2]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

[†]These authors have contributed equally to this work and share first authorship.

[‡]These authors share last authorship.

*** Correspondence:**
Katharina Lingelbach
Katharina.Lingelbach@iao.fraunhofer.de

**Keywords: Neuroadaptive Systems, Mental Effort Detection, Machine Learning, Multimodal Physiological Signals, Sensor Fusion, Ecological Validity, Neuroergonomics, Human-Machine Interaction**

Number of words: ~10,212/12,000

Number of Tables and Figures: 9/15

## Abstract

Humans' performance varies due to different internal and external states. These states affect the amount of mental resources available to successfully pursue a task. To prevent errors and incidents in human-machine-interaction (HMI), researchers must understand the intrinsic limitations of human cognitive information processing to enable the development of systems that can flexibly adapt to the users' currently available cognitive resources. Therefore, there is a pressing need for reliable and robust measures of a person's experienced mental effort that not only account for demands induced by the task itself but also consider situational and environmental influences and distractions. Since these measures should be applicable in real-world environments, they have to be developed and tested in realistic but still controlled experimental environments. We conducted a multimodal study with 18 participants (nine female, $M$ = 25.9 years, $SD$ = 3.8) of whom respiratory, ocular, cardiac, and brain activity (using functional near-infrared spectroscopy, fNIRS) were recorded during the execution of an adapted warship commander task with concurrent emotional speech distraction.

We introduce a machine learning architecture for multimodal (neuro-)physiological studies comprising feature engineering, model optimization, and model selection for an optimized combination of multimodal measurements in a cross-subject classification of experienced mental effort. We were able to reliably distinguish two different levels of mental effort when operationalized as experimentally induced task load from performance-based, neurophysiological, physiological, and visual data in a cross-subject classification. We consider this pipeline to be robust against noise, artifacts, and temporal sensor dropouts since it combines several sensor modalities. It is further, especially relevant for naturalistic applications where a precise evaluation of a good performance is difficult to obtain or not possible in the critical time window. When predicting subjectively perceived mental effort operationalized via self-reports, only performance-based measures allowed a successful prediction. Taken together, our tested multimodal classification approach contributes to the ecologically valid prediction of different states of mental effort and paves the way toward generalized state monitoring across individuals in realistic applications.

## 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 safety-critical decisions under high pressure. When designing technical systems to optimally support us with such tasks, an important goal persists in providing each individual user the opportunity to achieve the best possible performance outcomes with appropriate effort. Taking highly individualized demands and preferences into account, particularly in risky task settings, robust and reliable cross-subject classifications are crucial to avoid adverse consequences. Because real-life settings are characterized by multiple situational features, e.g., a potentially distracting social scene, extensive training, and testing of technical systems in ecologically valid experimental settings is a decisive prerequisite. For providing adequate reactions to a user's behavior, systems need sufficient input related to both performance and cognitive resource supply. While performance can be inspected via tracking the user's task-related progress, the actual pattern of invested cognitive resources should be derived from an enhanced scope of user-related information. These can emerge from advanced sensor technologies such as neurophysiological measurements, enabled by recent advances in portable neuroimaging techniques. Coupled with sophisticated signal processing and machine learning, these technical developments have paved the way for studying mental effort and its possible influences from a neuroergonomic perspective (von Lühmann, 2018; Charles and Nixon, 2019). Neuroergonomics aim to capture activation patterns in both – the central and peripheral nervous systems – associated with cognitive and emotional processes and thereby obtain insights into the relationship between neurophysiological functioning and behavioral outcomes in the context of work and in everyday life. In general, optimal neurophysiological functioning and behavioral outcomes are observed when the demands of a task match a person's capabilities (Bakker and Demerouti, 2007). Especially in performance-oriented contexts such as learning and training, safety-critical monitoring, or high-risk decision making, this fit between personal skills, abilities, and a task's requirements determines the quality of outcomes. However, a person's current performance is not constant, but it varies due to different physical conditions (e.g., illness, fatigue, or arousal), levels of experience and skills, subjective psychological states (e.g., stress, motivation, or emotions), and also external circumstances (e.g., noise, temperature, or distractions; see Hart & Staveland, 1988; Young et al., 2015). The characteristics of a task itself and its difficulty level also influence performance outcomes. A more difficult task requires more effort to be executed successfully. Further, concurrent distractions affect the allocation of attentional and cognitive resources (Lavie, 2010). According to the working memory model by Baddeley & Hitch (1974), it is thereby decisive whether distractions recruit resources from the same modality as the primary task. If both the primary task and the distracting stimulus compete for the same limited visual or auditory resources, performance in the primary task is more likely to decrease. However, based on an fMRI study, Sörqvist and colleagues (2016) argue for cognitive control mechanisms leading to decreased peripheral processing of task-irrelevant information under higher mental effort and, thus, reduced effects of distracting irrelevant stimuli. It might, therefore, be crucial how salient the distracting stimulus is. Salience describes the capacity of a stimulus to draw attention in a bottom-up, involuntary manner. Thus, a salient stimulus can disrupt goal-oriented, intentional, and top-down attention processes (Anikin, 2020). Previous studies found that irrelevant but intelligible speech showed disruptive effects on participants' performance in complex cognitive tasks (e.g., Banbury and Berry, 1998; Liebl et al., 2012). Intelligible speech might consequently increase the salience of a distractive stimulus. Other studies showed that the emotional intensity and the affective valence also influence the salience of a stimulus (e.g., Vuilleumier and Schwartz, 2001; Anikin, 2020). When a salient but task-irrelevant stimulus captures attention, people struggle to maintain goal-oriented behavior and such cognitive interference further impairs performance in the primary task (Dolcos et al., 2011; D'Andrea-Penna et al., 2017; Schweizer et al., 2019).

To ensure the maintenance of task-relevant cognitive processes, high-level cortical areas responsible for top-down regulation and executive functioning (mainly the prefrontal cortex; PFC) reduce task- or stimulus-irrelevant neural activities by inhibiting the processing of distractions (Klimesch, 2011). However, the salience of the distraction as well as the current effort required by the primary task might influence the effectiveness of the inhibitory mechanisms and successful maintenance of goal-directed processes. Research in mental effort, thus, has a long tradition in human factors and safety-critical applications (Hancock and Meshkati, 1988; Hancock and Desmond, 2001).

To quantify operators' or users' mental effort during the execution of a task, different measures are used that are 1) behavioral (i.e., performance-based) measures, 2) subjective assessments, and 3) neurophysiological measures (Paas et al., 2003; Chen et al., 2016; Zheng, 2017). Neurophysiological measures to investigate mental effort comprise brain activity, cardiac activity, blood pressure, respiration, and further skin-based and ocular measurements (see Tao et al., 2019, for a systematic review). Although each method has specific strengths and weaknesses, a neurophysiological approach has several decisive advantages, rendering it superior to performance-based measures or subjective assessments when used as the only measure. First, data can be obtained continuously in real-time without imposing an additional task (Babiloni, 2019). Moreover, it allows capturing not only task-related processes and effects but also correlates of inhibitory mechanisms (Klimesch, 2011). The interaction of multiple cognitive processes (task-specific and control mechanisms) has still to be fully unveiled. Therefore, the combination of multiple neurophysiological measures to capture both central and peripheral nervous system processes is of great advantage (Dirican and Göktürk, 2011; Debie et al., 2021). Furthermore, such multimodal approaches fusing data from several sources allow for a more comprehensive view of (neuro-)physiological processes related to mental effort (Uludağ and Roebroeck, 2014; Chen et al., 2016; Zheng, 2017; Wirzberger et al., 2018). By utilizing complementary measurement methods, one can compensate for specific disadvantages and combine the strength of each method, for example, regarding the temporal or spatial resolution (Uludağ and Roebroeck, 2014; Zhang et al., 2020; Debie et al., 2021).

A major challenge for multimodal approaches is the fusion of data. Machine learning (ML) methods provide solutions to compare and combine data streams from different measurements. ML algorithms are becoming increasingly popular in computational neuroscience (Lemm et al., 2011; Vu et al., 2018). 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. Suppose the model "learns" the relevant characteristics in the neurophysiological data allowing to distinguish between high and low perceived mental effort during the training phase. In that case, the model can later predict a person's perceived mental effort by using the same informative features in a test phase. This data-driven approach allows us to exploratorily identify patterns that are informative features regarding the current mental effort in a multimodal dataset (see also Herms et al., 2018).

Particularly in data-driven research paradigms, most research has been done using within-subject models, which tend to perform better than cross-subject models which face the challenge of high inter-individual variability in physiological signals (Waytowich et al., 2016). However, there are already some attempts to train cross-subject models that overcome the need for subject-specific information during training (Lawhern et al., 2018; Lyu et al., 2021). Still, the subject-independent classification is a major challenge in the field (Lotte et al., 2018) and a solution to this problem is crucial for the development of "plug and play" real-time state recognition systems (Liu et al., 2019, 251) as well as the resource-conserving exploitation of already available large datasets without time-consuming individual calibration sessions. To achieve significant subject-independent classifications,

which allow conclusions to be drawn on real-world applications, it is of utmost importance to adapt experimental designs regarding their correspondence to the real world. Especially in cognitive neurosciences, researchers build their conclusions on findings acquired in laboratory-based experiments. The standardized experimental procedures in controlled laboratory environments minimize the influences of confounding variables. While this standard procedure provides high internal validity and allows for systematic testing of specific hypotheses and conclusions regarding causal or correlative relationships between the experimental variables and manipulations, it often lacks generalization when transferring findings from the somewhat artificial laboratory to a natural, non-experimental setting (Orne, 1962; Neisser, 1976; Hoc, 2001). Thus, researchers have a high interest in expanding their methods by experimental designs with more naturalistic characteristics (Ladouce et al., 2017) to account for the complex interactions and interdependencies between cognition and the environmental context. With the increasing availability of portable devices providing high-quality data, researchers have the opportunity to acquire neurophysiological data in more dynamic and complex environments (Ladouce et al., 2017).

In recent neuroergonomic research, functional near-infrared spectroscopy (fNIRS) has been used to measure changes in cortical hemodynamic concentration during a cognitive performance (primary workload) and emotional tasks with high ecological validity (Curtin and Ayaz, 2018; von Lühmann, 2018; e.g., Benerradi et al., 2019; Midha et al., 2021). fNIRS is an optical imaging technology allowing researchers to measure local oxy-hemoglobin (HbO) and deoxy-hemoglobin (HbR) changes in cortical regions. Higher mental effort is associated with a hemodynamic response in the prefrontal cortex (Izzetoglu et al., 2003; PFC; Ayaz et al., 2012; Herff et al., 2014). The PFC is crucial for executive functions like maintaining goal-directed behavior as well as the suppression of distractions (Miller et al., 2002; Dehais et al., 2020). In addition to changes in the central nervous system, an increased mental effort also leads to adaptation processes in the autonomic nervous system. The autonomic nervous system, as part of the peripheral nervous system, regulates involuntary (automatic) physiologic processes to maintain homeostasis in the bodily functioning (Babiloni, 2019; Matthews et al., 2005). Increased mental effort is associated with a decrease in parasympathetic nervous system activity and increased sympathetic nervous system activity (Wierwille, 1979; Kramer, 1991; Backs, 2000). Typical correlates of the autonomic nervous system 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 (e.g., Kramer, 1991; Dirican and Göktürk, 2011; Dan and Reiner, 2017; Charles and Nixon, 2019; Tao et al., 2019; Romine et al., 2020). Especially in visually demanding tasks, the increased effort is reliably associated with lower blink rates, increased fixation duration (Charles and Nixon, 2019), and increased pupil diameter (Hosseini et al., 2017; Appel et al., 2018). However, ocular measures are also affected by emotional stimuli (Bradley et al., 2008). Similarly, heart rate is positively correlated with activation and arousal (Birbaumer and Schmidt, 2010) that might also be elicited by cognitive demands, engagement, and mental effort, and a decrease in heart rate variability is considered a reliable indicator of increased mental effort (De Rivecourt et al., 2008; Durantin et al., 2014; Charles and Nixon, 2019). Besides, heart rate is also highly susceptible to emotional arousal (Brouwer et al., 2015; Schneider et al., 2019), e.g., due to frustration and anger. It might thus be an appropriate measure of mental states of stress and increased tension (Taelman et al., 2009). Peripheral physiological measures provide an inexpensive, unintrusive, and robust measure of sympathetic nervous system activity. They are suitable as an additional method for verifying mental effort levels.

Taken together, in the present study, we investigate mental effort which refers to the cognitive resources that are allocated to meet both subjective and objective performance criteria imposed by a

task (see Paas et al., 2003, for a definition and differentiation of the term). High mental effort occurs when the demands induced by the task as well as situational characteristics exceed the available resources to prioritize and pursue a specific goal. This state often leads to exhaustion, stress, fatigue, and consequently insufficient time and energy for adequate information processing or resource allocation towards goal-directed processes (Gevins and Smith, 2003; e.g., Bowling et al., 2015; Young et al., 2015), which can then be detrimental to performance and lead to errors and incidents. Therefore, we used a quasi-realistic experimental 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 (Becker et al., 2021). It, thus, comprises a complex sequence of cognitive operations and does not focus on individual cognitive performance areas. Therefore, the task provides a good representation of the requirements of real-world monitoring tasks in a working scenario. We used auditory speech-based stimuli to account for concurrent distractions (Burkhardt et al., 2005). To explore the interaction of emotional processing and mental effort, we used three emotional types of utterances with neutrally, positively, and negatively emotionally associated prosody. During the execution of the experimental task, we recorded multiple neurophysiological measures. We were particularly interested in finding an optimized multimodal machine learning estimation capable of predicting experienced mental effort induced by the task itself but also by the suppression of situational auditory distractions in a complex close-to-realistic environment. When referring to neurophysiological measures, we mean both brain-related as well as peripheral physiological activity.

We explored different data fusion and classifying strategies for a cross-subject prediction of experienced mental effort in a complex close-to-realistic environment based on multimodal neurophysiological and behavioral data. First, we investigated whether a combined prediction of various ML models is superior to the prediction of a single model (RQ1). Further, we explored whether a multimodal classification that combines the predictions of different modalities is superior to a unimodal prediction (RQ2).

## Materials and Methods

### Participants

Volunteers for participation filled in a screening questionnaire that checks eligibility for study participation and collects demographic characteristics. Individuals with insufficient knowledge of the German language or limited color vision were not admitted to the study because these abilities are required for the experimental procedure. 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); all had a normal or corrected-to-normal vision. They received monetary compensation for their voluntary participation. Before their participation, they signed an informed consent according to the recommendations of the Declaration of Helsinki. The study was approved by the ethics committee of the Medical Faculty of the University of Tuebingen, Germany (ID: 827/2020BO1).

## Experimental Task

Participants had to perform an adapted version of a warship commander task (WCT, Pacific Science & Engineering Group, 2003; adapted by Becker et al., 2021). The WCT is a quasi-realistic, complex navy command and control task designed as a basic analog to a Navy air warfare task, suitable to investigate various cognitive processes of human decision-making and action execution (St John et al., 2003). Here, we used a non-military and safety-critical task. Participants had to identify two different flying objects on a simulated radar screen around an airport: registered drones (neutral, non-critical objects), or non-registered (critical) drones. Thus, they had to prevent the non-registered drones, potentially being a safety issue, from entering the airport's air space. More specifically, 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. (2021) for a more detailed description of the scoring system, and see Figure 1 for an overview of the interface.

[Insert Figure 1 here]

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; Burkhardt et al., 2005). 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.

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 (operationalized with the difficulty level) was alternated across blocks. Half of the subjects started with high task load and the other half with low task load. Similarly, the concurrent emotional condition (operationalized with different auditory distractions) was randomized 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. They had to track 36 objects in the high task load condition, 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.

[Insert Figure 2 here]

## Questionnaires

Subjectively perceived mental effort and affective state were assessed during each scenario. We asked participants to rate the effort and frustration they perceived in the previous block on the corresponding NASA TLX effort and frustration subscales (Hart and Staveland, 1988) after each experimental block. In addition, we asked participants to evaluate their affective state using the dimensional EmojiGrid (Toet et al., 2018) and the categorical Circumplex Affect Assessment Tool (CAAT; Cardoso et al., 2013).

After the experiment, participants answered questionnaires regarding personal traits that might have influenced their performance and behavior during the study. Therefore, we used the short version of the German Big Five Inventory (BFI-K) to capture personality characteristics (Rammstedt and John, 2005) and the German State-Trait-Anxiety Inventory (STAI; Laux et al., 1981). Furthermore, we asked about the general ability to concentrate, focus, and not get distracted using the Attention and Performance Self-Assessment (APSA; Bankstahl and Görtelmeyer, 2013) and the German language version of the Barratt Impulsiveness Scale-11 (BIS; Hartmann et al., 2011).

Here, we only used the NASA TLX ratings for the labeling of the (neuro-)physiological data for the ML classification, while the other subjective measures were not of interest in this analysis.

## Data Collection: 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 with a sampling rate of 60Hz.

To capture changes in physiological responses, participants were wearing a BioHarness™ belt recording electrocardiographic (ECG), respiration, and temperature signals with a sampling rate of 1 Hz. Here, we used automatically computed, aggregated scores for the heart rate, the heart rate variability, and respiration rate and amplitude from the device.

Participants' brain activity was recorded with a NIRx NIRSport2 system which emits light at two wavelengths, 760 and 850 nm. Aurora fNIRS Recording Software with a sampling rate of 5.8 Hz was used for data collection. To capture regions associated with mental effort, the 14 source optodes and 14 detector optodes were placed at the prefrontal cortex (Ayaz et al., 2012; Scheunemann et al., 2019) using the fNIRS Optodes' Location Decider (fOLD) toolbox (Zimeo Morais et al., 2018). See Figure 3 for the complete montage and the resulting 41 channels.

[Insert Figure 3 here]

## Data Preprocessing and Machine Learning

All data pre-processing and machine learning analyses steps were performed with custom-written scripts in *R* and python™. Every continuous raw data stream was cut into non-overlapping 60-second intervals beginning from the start of each experimental trial (see Figure 2).

Before feeding the data into the classification procedure, we applied the following data cleaning and preprocessing steps per modality as explained in detail next.

## Preprocessing of Eye-Tracking Data

First, we cleaned the continuous eye tracker data using the *eyetrackingR* package in *R* (Dink and Ferguson, 2015). Missing values were filled using linear interpolation. We extracted 855 trials with a length of 60 seconds (on average 47.5 trials per subject, $SD = 0.9$). Next, we used the validity index provided by the eye tracker itself to remove non-consistent data segments from further analysis. The index 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 pre-processing of the pupil dilation data, we used the *PupillometryR R*-package (Forbes, 2020). First, we calculated a simple linear regression of one pupil against the other and vice versa,

per subject and trial to smooth out small artifacts (Jackson and Sirois, 2009). Afterward, we computed the mean of both pupils and filtered the data using the median of a rolling window with a size of 11 samples (default). To control for the variance of pupil sizes between participants, we applied a subject-wise *z*-score normalization of the pupil dilation.

For the computation of fixations, we used the *saccades R*-package (von der Malsburg, 2015). 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 for the number and the duration of fixations separately for each subject.

### Preprocessing of Physiological Data

The epoching in non-overlapping time windows of 60-second intervals of the raw data from the BioHarness™ 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 normalization.

### Preprocessing of fNIRS Data

For the pre-processing of the fNIRS data, we used the libraries *MNE-Python* (Gramfort et al., 2014) and its extension *MNE-NIRS* (Luke et al., 2021). We followed the guidelines as presented by Yücel et al. (2021), including channel pruning, detrending, removal of signal noise, and correction of artifacts. 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 (Pollonini et al., 2014). Channels with a scalp-coupling index below 0.5 were marked as bad channels. We further applied a temporal derivative distribution repair to account for a baseline shift and spike artifacts (Fishburn et al., 2019). Channels marked as bad were interpolated, with the nearest channel providing good data quality. Afterward, a short-separation regression was used, subtracting short-channel data from the standard long-channel signal to remove signal changes in the extracerebral layer and correct for systemic signals contaminating the brain activity measured in the long-channel (Saager and Berger, 2005; Yücel et al., 2021). Next, the modified Beer-Lambert Law was applied to transform the optical density data into oxygenated (HbO) and deoxygenated (HbR) hemoglobin levels (Beer, 1852). As suggested by the standard fNIRS MATLAB application *Homer3* (Huppert et al., 2009), we used a partial pathlength factor of 6 for the transformation. Further, data were filtered using a fourth-order zero-phase Butterworth bandpass filter with cutoff frequencies of 0.05 and 0.7 Hz and with a transition bandwidth of 0.02 and 0.2 Hz to remove instrumental and physiological noise (such as heartbeat and respiration). The resulting data was cut into epochs with a length of 60 seconds. We then applied a channel-wise *z*-score normalization. In total, 730 trials were left for further analyses (on average 40.6 trials per subject, *SD* = 9.6).

### Feature Extraction

Our feature space, thus, comprised the aggregated brain activity (fNIRS), physiological, ocular and performance-related measures. Table 1 gives an overview of the included features for each modality. From the computed fixations, we used the number of fixations, the total duration, the average duration, and the standard deviation of the duration of fixations per subject and trial as features for the classification. From the pupil dilation data, we calculated the mean, standard deviation, skewness, and kurtosis per subject and trial as features for the classification based on the cleaned data. From the physiological data, we computed the mean, standard deviation, skewness, and kurtosis per subject

and trial of the heart rate, heart rate variability, respiration rate, respiratory amplitude, and body temperature. From the fNIRS data, we extracted the mean, the standard deviation, the peak-to-peak (PTP) amplitude, the skewness, and kurtosis of the hemodynamic response as features with the *mne-features* package (Schiratti et al., 2018). Furthermore, we added behavioral measures that capture the participant's performance (average reaction time and cumulative accuracy per trial).

[Insert Table 1 here]

## Ground Truth for Machine Learning

Our main goal was to identify the mental effort experienced by an individual across subjects using machine learning (e.g., Keles et al., 2021; Minkley et al., 2021). Since experimentally manipulated task load was further influenced by situational demands (e.g., inhibiting task-irrelevant auditory emotional distraction), the actually perceived mental effort might not be fully captured by the experimental condition. Therefore, we explored two approaches to operationalize 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 categorized values above the threshold as "high mental effort" and other values 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.

For comparison reasons, we also 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. By introducing this split of the data, we wanted to be able to not only find measures that can distinguish between low and high mental effort but also identify predictive measures for very high perceived mental effort potentially reflecting cognitive overload. With this 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 the perceived mental effort in the different experimental conditions. Hereby we wanted to control for confounding effects that are typical for self-reports, e.g., consistency effects or social desirability effects.

## 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, Pedregosa et al., 2011).

[Insert Figure 4 here]

Figure 4 shows a schematic representation of our multimodal classification scheme and cross-subject validation procedure using multiple randomized grid search operations. 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, we wanted to 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 generalize across individuals and allows to exploit datasets already collected in a similar context for the training of the models.

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*-standardized (see *Standard Scaling* per modality in Figure 4). Standard scaling ensured, that for each feature the mean is zero and standard deviation is one and, 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. The hyperparameters for each classifier were optimized by means of a cross-validated randomized grid search with a maximum number of 100 iterations per cross-validation and the 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 might hence be a more computational efficient approach. This is especially important when transferring the findings to real-world scenarios. Due to our cross-subject approach, the selected hyperparameters varied for each predicted test subject.

Afterward, we combined these classifiers using a voting classifier implemented in the *mlxtend* package (version 0.19.0, Raschka, 2018). The ensemble classifier makes predictions based on aggregating the findings 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). This requires that there is a sufficient number of weak learners, and these are sufficiently diverse (as is the case with our classifier selection above). Further, 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 again optimized using a cross-validated randomized grid search with a maximum number of 100 iterations. With this procedure, we were able to compare the predictions of the single classifiers to a weighted combination of these classifiers. This way, we were able to find a suitable approach for every modality (brain activity, physiological activity, ocular measures, and performance).

For a multimodal approach, we entered the resulting voting classifiers for the four different modalities into a second voting classifier that combines the unimodal predictions into a final multimodal prediction of subjectively perceived mental effort. This second voting classifier also assigned weights to the different modality-specific classifiers. We again optimized these weights and the voting procedure (soft or hard) using a cross-validated randomized grid search with a maximum number of 100 iterations.

We report the average $F_1$ score and a confusion matrix in the training set and the test subject to evaluate the models' performances. The $F_1$ score can be interpreted as a weighted average or "harmonic mean" of precision and recall. Precision refers to the number of samples predicted as positive that are actually 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 minimizing false positives – and hence is also our target metric for the cross-validation. This means that the model parameters are optimized to achieve a high $F_1$ score as a score of 1 indicates a very good model performance, and a value of 0 a bad model performance. Still, in our scenario, the recall (the fraction of "high mental effort" cases that were actually detected by the classifiers) is the most important metric to assess the usefulness and the practical applicability of the models, since robust identification of high mental effort states potentially allows to prevent safety-critical situations and promote sustainable working conditions by offering assistance.

## Results

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

### Unimodal Predictions

The performance of the different classification approaches as assessed by the $F_1$ score is visualized via boxplots; see Figure 5 as an example. Significant differences can be discerned from the non-overlapping notches of the respective boxes, which mark the upper and lower boundaries of bootstrapped 95% confidence intervals (CI). The upper limit of a dummy classifier which only considers the distribution of the outcome classes for its prediction and, hence, represents an empirical chance level estimate, is plotted as a continuous dashed grey line in all the subplots of Figure 5. For a prediction to be significantly better than chance (significance level below .05), its bootstrapped mean must not overlap with this grey line. For a significance level below .01, the lower boundaries of the CI can be used (Cumming and Finch, 2005).

Based on the $F_1$ scores, we do not see substantially better performance when using a larger validation set of two subjects, neither for the median split (see Figure 5 and Supplementary Figure 3) nor for the upper quartile split (see Supplementary Figures 3, 7, and 11) or the prediction of the experimentally induced task load (see 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.

[Insert Figure 5 here]

Figure 5 shows the performance of the different classifiers for predicting low and high mental effort in a median-split-based unimodal (Figure 5A, B, D, E) as well as multimodal approach (Figure 5C; elaborated on in Section Multimodal Predictions). Regarding the unimodal classifications, we see a better prediction of the subjectively perceived mental effort based on performance data (Figure 5E) than based on ocular, physiological, or brain activity measures (Figures 5A, B, and D). Based on the latter measures, none of the unimodal classifications perform significantly better than the dummy classifier (Figure 5F) 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 randomized grid search. See Figure 6 for an overview of the selected voting procedures and the allocated weights per modality.

[Insert Figure 6 here]

First, most of the unimodal voting classifiers are overfitted, as indicated by the large deviation between training and test performance, except for the one based on performance data (see Figures 5A, B, D, and E). Accordingly, we see the best performance for classifying subjectively perceived mental effort based on performance. 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 outliers with low classification performance, which represent 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 generalization over subjects is much more challenging when using the neurophysiological compared to the performance-based features.

**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. Regarding the $F_1$ score, we see strong overfitting (see 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 (see 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. When it comes to this comparison based on brain activity (see Figure 7A), we see that both distributions indicate that we were not able to derive good predictions from the unimodal classification. Especially the voting classifier has a high number of falsely identified "High Mental Effort" cases (False Positives), leading to a recall of 45.6% and a precision of only 39.3%. The best performing single classifiers have an average recall of 57.5% and an average precision of 49.8%.

[Insert Figure 7 here]

**Unimodal Predictions – Physiological measures**

For classifying subjectively perceived mental effort based on physiological measures such as heart rate, respiration, and body temperature, half of the test subjects had voting classifiers using soft voting, and half had ones with hard voting (see Figure 6B). The weighting of the classifiers varied considerably, with the KNN obtaining the highest average weights. The voting classifier (see Figure 5B) showed strong overfitting, and its performance in the test subject was neither significantly better than any of the single classifiers nor a dummy classifier. Regarding the percentage of correctly and falsely classified cases (see 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, which 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%).

**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 (see Figure 6C). KNN and SVM were weighted highest on average. The $F_1$ score of the voting classifier (see Figure 5D) did not show a significantly better classification performance than the dummy classifier. Based on the percentage of correctly and falsely classified cases (see Figure 7D), the pattern 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%).

**Unimodal Predictions – Performance**

At last, we predicted subjectively perceived mental effort based on performance data (accuracy and speed). 27.8% of the test subjects had voting classifiers using soft voting, and 72.2% used hard voting (see Figure 6D). The SVM was 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. Based on the percentage of correctly and falsely classified cases (see Figure 7E), the colors indicate a better performance compared to the confusion matrices for brain, physiological and ocular activity. The voting classifier again had a high number of 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%.

**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 to the median-split-based results, we observed decreased classifiers’ performance even below dummy classifier performance (see Supplementary Figures 7). This might be explained by the fact that we reframed a binary prediction problem with evenly distributed classed 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 much harder (reflected in the recall; see Supplementary Figures 9).

**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 operationalized 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. The best unimodal voting classifications were obtained based on performance measures. Interestingly, other classification models were favored in the

unimodal voting, and the distribution between soft- and hard voting differed compared to the subjectively based approach, with soft voting being used more often (see Supplementary Figure 17).

### Multimodal Predictions based on the Median Split

In a final step, we combined the different modalities into a multimodal prediction, exploiting all measurements' information. See Figure 5C and Figure 7C for the performance of the multimodal voting classifier, and see Figure 8A for the average allocated weights to the different modalities.

[Insert Figure 8 here]

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 the performance measures to predict subjectively perceived mental effort, thereby turning it into a unimodal classifier. Based on the $F_1$ score, the voting classifier led to a significantly better classification than the dummy classifier (see Figure 5C). Based on the percentage of correctly and falsely classified cases (see Figure 7C), the multimodal classifier naturally showed the same performance as the classifier based on performance data, with an average recall of 78.5% and an average precision of 57.6%.

To explore how well the multimodal classifier would perform without the information of the performance measures (speed and accuracy), we restricted the multimodal classifier only to use neurophysiological and visual measures. This approach is especially relevant for naturalistic applications where a precise evaluation of a good performance is difficult to obtain or not possible in the critical time window. For the multimodal prediction without performance, brain activity was weighted highest (see Figure 8B). However, classifiers revealed strong overfitting during the training, and the average performance was decreased to a chance level (average recall: 40.6% and average precision: 38.3%; see Figure 5C).

### Multimodal Predictions based on the Upper Quartile Split

With the upper quartile split, we had a fundamentally different allocation of weights. Brain and ocular activity were assigned the high weights (see Figure 8B) and performance obtained only very small weights. Accordingly, the exclusion of performance-based measures had almost no effect on the allocation of weights (see Supplementary Figure 10B) as well as on the average performance of the multimodal classifiers (see Supplementary Figure 7C). In two out of eighteen test subjects, the multimodal classification revealed the highest performance indicating that in these cases, the exploiting of the combined information was beneficial (see Supplementary Table 2). However, on average, the multimodal classification based on an upper quartile split was not superior to the unimodal classifiers as assessed by the $F_1$ score and did not perform significantly better than the dummy classifier (average recall: 21.9% and average precision: 18.5%; see Supplementary Figure 7).

### 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 for the prediction of the experimentally induced task load solely used the performance measures. The average prediction performance was very high and significantly better than the one of a dummy classifier (average recall: 99.7% and average precision: 91.3%; see Supplementary Figure 19A). When we only allowed neurophysiological and visual measures as features, visual measures were weighted highest (see Supplementary Figure 19B). In this case, the

603 average performance of the multimodal classifiers was also significantly above the chance level, with
604 an average recall of 82.7% and an average precision of 58.7%, indicating a successful identification
605 of mental effort based on neurophysiological, physiological, and visual measures (see Supplementary
606 Figure 16C).

607 **Discussion**

608 The purpose of our study was to explore and compare predictive correlates of the subjectively
609 experienced and experimentally induced mental effort in a complex close-to-realistic environment
610 using a cross-subject multimodal machine learning approach. This is important because a robust
611 method to detect high mental effort in real-time and under real-world conditions is needed to
612 facilitate the adaptation of systems to users' current mental resources. These adaptive systems could
613 enable people to perform at their best due to an optimal level of demand. Especially in tasks with a
614 high security risk, system engineers should strive to ensure that people are neither overburdened nor
615 bored and, thus, prone to mistakes.

616 In the here presented study, we obtained multimodal data from participants performing a quasi-
617 realistic experimental task with varying task complexity and salient distractions. In our analyses, we
618 fitted a complex cross-subject multilevel classification and tested its performance with a leave-one-
619 out approach. We provided insights into which classifiers perform best in which modality and further
620 explored whether combinations of ML models are superior to predictions of single models. In a
621 second step, we combined the different unimodal classifiers into a multimodal voting classifier. We
622 hence investigated whether this multimodal classification is superior to a unimodal prediction.

623 For each modality, we found a different set of classifiers that were better performing in the prediction
624 and, therefore, also considered more informative in the unimodal voting.

625 When predicting subjectively perceived mental effort, LDA and LR performed best and were
626 weighted highest in the classifications based on brain activity. Whereas, in physiological activity, the
627 highest weights were assigned to KNN, RFC, and SVM. Regarding visual activity, the GNB had the
628 best average performance among the test subjects – together with the KNN in the median split
629 prediction. These models aiming to predict subjectively perceived mental effort based on brain
630 activity, physiological activity, and visual measures were strongly overfitted, and their performances
631 in the test subjects were not significantly better than the dummy classifier. In performance-related
632 measures, the GNB, RFC, and SVM performed significantly better than the dummy classifier when
633 predicting subjectively mental effort based on a median split. Using the upper quartile split, the KNN
634 and SVM showed the highest, but still, chance-level-like performances. When predicting
635 experimentally induced mental effort using brain activity, GNB, KNN, and SVM performed above
636 chance level and were assigned the highest weights in the unimodal voting. For the physiological
637 activity, all classifiers except the KNN reached above chance level performance, with the highest
638 average weight in the unimodal voting being assigned to the SVM. For visual and performance
639 measures, we did not see substantial differences between the classification performance of the single
640 models, with all performing above chance level.

641 Regarding the unimodal predictions of subjectively perceived mental effort, we see that a weighted
642 combination of classifiers (LR, LDA, GNB, KNN, RFC, and SVM) was not superior to the single
643 classifiers neither when using the median nor the upper quartile split. When we combined the
644 different modalities to a joined prediction of subjectively perceived mental effort, only the
645 performance modality was taken into account. Hence, our multimodal classification must rather be

considered a unimodal (performance-based) prediction. Removing the performance information from the multimodal voting classifier in a second step led to a considerable drop below the chance level in the classifiers' average performance. When investigating the upper quartile split classification, performance was less predictive in identifying cases of exceptionally high perceived mental effort or potentially even "cognitive overload". Multimodal voting classifiers trained to identify these high mental effort states operationalized by an upper quartile split gave higher weights to neurophysiological and visual measures compared to performance measures. This implies 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. However, these classifiers revealed relatively poor performance due to the challenge of a highly imbalanced classification problem (i.e., outlier detection problem). Based on the results obtained, there is, thus, still a great need for research on suitable features and algorithms for the robust prediction of current mental states in naturalistic scenarios. In summary, in a unimodal voting approach, subjectively perceived mental effort could best be predicted using performance-based measures. Other modalities or multimodal approaches were not superior to this classification.

For the classification of the experimentally induced task load, all modalities were able to predict mental effort with high performances already in certain single classifier models. A unimodal weighted combination of these classifiers was not superior to the single classifiers in any modality. The performance modality was also most predictive of task load. Consequently, the multimodal classifier again transformed into unimodal voting, since it solely relied on the performance modality. When excluding the performance-based features, a multimodal prediction based on neurophysiological, physiological, and ocular activity was still significantly above the chance level estimated by a dummy classifier.

This indicates that it is possible to distinguish different mental effort states operationalized as experimentally induced task load based on neurophysiological and visual data acquired in a close-to-realistic environment and cross-subject classification approach. However, it was not possible to replicate these results for subjectively perceived mental effort. Deviations between these two ground truth approaches might be explained by the retrospective nature of self-reports implying an automatic evaluation process as they depend on the individual's perception, reasoning, and unverifiable introspection (Ranchet et al., 2017). They are, therefore, vulnerable to various perceptual and response biases like social desirability (Dirican and Göktürk, 2011; Matthews et al., 2020). These post-hoc processes might not be reflected in and could be learned from neurophysiological and physiological measures during the task itself.

For all classification approaches, we observed substantial variation in the performance of classifiers between the test subjects. Some subjects had $F_1$ scores above 0.8 (see Supplementary Table 1). Other subjects deviated strongly in their (neuro-)physiological reactions and, hence, did not fit well into the patterns identified among the subjects in the training set. These results are in line with the findings by Causse colleagues (2017), who concluded that it is quite challenging to identify mental states based on hemodynamic 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 (Allison and Neuper, 2010). Our results highlight the need for suitable methods a) to identify subjects who are potentially difficult to predict due to their heterogeneity compared to the training set or other inter-individual differences, and b) to

facilitate transfer learning for these individuals as well; for instance, by standardizing and transforming correlates into a common feature space (de Cheveigné et al., 2019).

## Limitations and Future Research

We consider some aspects of this study as objects for further improvement. First, we used a quite homogeneous sample regarding the socio-demographic characteristics with young age and high education, limiting our results' generalizability. Although one intuitive idea might be to increase the sample size, there is some debate about the relationship between the sample size and the heterogeneity in the sample in the ML community. As Cearns and colleagues (2019) described, although machine learning classifiers might (surprisingly) perform exceptionally well in relatively small datasets, one could then suddenly observe a decrease in accuracy with the addition of some more samples. This is probably because of an increasing heterogeneity after profiting from a relatively small and homogenous dataset before. 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 generalizability and attain good prediction accuracy on an individual-level (Bzdok and Meyer-Lindenberg, 2018; Dwyer et al., 2018). Nevertheless, others propose that there are ways to reach better generalizability, less bias, and enable ML models to perform well – even on a subject-level in small datasets. Orrù et al. (2020), for example, suggest the use of simple classifiers or ensemble learning methods instead of complex neural networks. Cearns and colleagues (2019) highlight the importance of suitable cross-validation methods. Especially in the case of physiological datasets comprising large differences between subjects, 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 behavior or neurophysiological reactions (Dwyer et al., 2018). One interesting idea to address this problem is data augmentation (Lashgari et al., 2020; Bird et al., 2021). Data augmentation means generating new samples by transforming existing samples to extend a dataset. For example, by the use of 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 (Zanini and Colombini, 2020). Another suggested method to improve generalizability across subjects might be multiway canonical correlation analysis (MCCA). An approach that allows to combine multiple data sets into a common representation and, thereby, achieves the denoising of data, and dimensionality reduction, based on shared components across subjects (de Cheveigné et al., 2019). In general, a shared goal of researchers should be to make data sets more comparable and maybe even combinable (Abrams et al., 2021). Large, combined datasets may then allow the identification of predictive patterns that are generalizable to a variety of people and situations.

Second, one should aim for high data quality and richness of annotation. Further artifact analyses, which can, for example, correct distortions induced by movement (von Lühmann et al., 2019), or the implementation of inclusion criteria on the subject-, trial-, and channel-level could be explored in order to improve poor signal-to-noise ratios. Friedman and colleagues (2019), who used an XGBoost classifier on EEG data, applied some extensive and rigorous 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 cf., Unni et al., 2017). In addition, tasks eliciting significantly different performance and subjects showing unusual behavior were also excluded from their analysis. 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 standardized evaluation methods to decide which preprocessing step is beneficial and adequate.

A third limitation pertains to the montage of the fNIRS optodes. Based on previous research (e.g., Ayaz et al., 2012), 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 (Unni et al., 2017). Integrating these regions would have allowed identifying features for the classification from larger functional networks that might play a crucial role in distinguishing mental states and cognitive control mechanisms (Sörqvist et al., 2016; García-Pacios et al., 2017). Increased activity in the frontoparietal network is, for example, associated with task-related working memory (WM) processes (e.g., Curtis, 2006; Martínez-Vázquez and Gail, 2018), whereas increased connectivity between frontal and sensory areas are linked to the suppression of distractors (García-Pacios et al., 2017).

## Feature Selection and Data Fusion in Machine Learning

A crucial aim of this study was the selection and fusion of informative sources for robust 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 behavioral measures of performance (accuracy and speed). However, this selection was naturally not exhaustive. Other measures, such as electroencephalography or electrodermal activity (e.g., Vanneste et al., 2021), could provide useful information about cognitive and physiological processes related to mental effort. In addition, one could also explore more behavior-related measures such as speech (e.g., Yap et al., 2015) or gaze (e.g., Marquart et al., 2015). These measures might 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 and colleagues (2019) investigated considerably more elaborated features like connectivity and complexity metrics which might capture helpful information regarding relationships within and between neuronal networks. Further investigations are needed regarding the predictive quality of such aggregated features. Furthermore, the additional value of feature selection and wrapping methods to reduce the complexity of the feature space without losing predictive information should be explored (Gottemukkula and Derakhshani, 2011; Aydin, 2020). Such dimensionality reduction (e.g., sequential feature forward selection, principal component analysis (PCA), or linear discriminant analysis (LDA), might be a way to improve classifiers' performance by keeping only the most informative aspects in our data and reducing noise. 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 (e.g., Lawhern et al., 2018; Chakraborty et al., 2019; Asgher et al., 2020). 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

fulfill in naturalistic settings with multimodal measurement methods using different measurement devices.

After defining the feature space, a strategy on the classification level for selecting, merging, combining, and weighting several classifier models and modalities must be chosen. Especially in multimodal datasets, the optimal timepoint for different modalities fusion must be found. Hereby, one has to balance the computational power as well as the size of the required dataset against the benefit of a fine-tuned combination of optimally stacked or voted classifiers. Here, the investigation of early and late fusion approaches as used in the field of robotics might be helpful. While early fusion promotes an early combination of all data points and fitting classifiers to multidimensional data, late fusion pertains to a more fine-grained pipeline where several classifiers are fit to proportions of the dataset and 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 second step, we combined these different modalities into a joint prediction. Exploring early and late fusion strategies is especially important when one wants to account for temporal dynamics in the different measures or the realization of real-time mental state monitoring. The review of Debie and colleagues (2021) provides a comprehensive overview of the different fusion stages when identifying mental effort based on neurophysiological measures.

## Practical Implications and Conclusion

Our proposed multimodal classification approach contributes to the ecologically valid distinction and identification of different states of mental effort. It paves the way toward generalized state monitoring across individuals in realistic applications.

Interestingly, the choice of ground truth had a fundamental influence on the classification performance. Subjectively perceived mental effort operationalized via self-reports could only be predicted from performance-based measures. The experimentally induced task load could be predicted based on performance-based measures but also based on neurophysiological and visual measures. This classification pipeline can, thus, be considered quite robust against noise, artifacts, and temporal sensor dropouts for mental effort predictions since it combines several sensor modalities.

Our results allow researchers and practitioners to select appropriate methods for their research questions or application scenarios, taking into account limited resources or constraints imposed by the environment. 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-centered applications such as adaptive assistance systems.

## References


Abrams, M. B., Bjaalie, J. G., Das, S., Egan, G. F., Ghosh, S. S., Goscinski, W. J., et al. (2021). A standards organization for open and FAIR neuroscience: The international neuroinformatics coordinating facility. *Neuroinformatics*. doi: 10.1007/s12021-020-09509-0.

Allison, B. Z., and Neuper, C. (2010). "Could anyone use a BCI?," in *Brain-Computer Interfaces: Applying our Minds to Human-Computer Interaction*, eds. D. S. Tan and A. Nijholt (London: Springer), 35–54. doi: 10.1007/978-1-84996-272-8_3.

Anikin, A. (2020). The link between auditory salience and emotion intensity. *Cognition and Emotion* 34, 1246–1259. doi: 10.1080/02699931.2020.1736992.

Appel, T., Scharinger, C., Gerjets, P., and Kasneci, E. (2018). Cross-subject workload classification using pupil-related measures. in *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, eds. B. Sharif and K. Krejtz (Warsaw Poland: Association for Computing Machinery), 1–8. doi: 10.1145/3204493.3204531.

Asgher, U., Khalil, K., Khan, M. J., Ahmad, R., Butt, S. I., Ayaz, Y., et al. (2020). Enhanced accuracy for multiclass mental workload detection using long short-term memory for brain–computer interface. *Frontiers in Neuroscience* 14, 584. doi: 10.3389/fnins.2020.00584.

Ayaz, H., Shewokis, P. A., Bunce, S., Izzetoglu, K., Willems, B., and Onaral, B. (2012). Optical brain monitoring for operator training and mental workload assessment. *NeuroImage* 59, 36–47. doi: 10.1016/j.neuroimage.2011.06.023.

Aydin, E. A. (2020). Subject-Specific feature selection for near infrared spectroscopy based brain-computer interfaces. *Computer Methods and Programs in Biomedicine* 195, 105535. doi: 10.1016/j.cmpb.2020.105535.

Babiloni, F. (2019). Mental workload monitoring: New perspectives from neuroscience. in Communications in Computer and Information Science., eds. L. Longo and M. C. Leva (Cham: Springer International Publishing), 3–19. doi: 10.1007/978-3-030-32423-0_1.

Backs, R. W. (2000). Application of psychophysiological models to mental workload. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 44, 464–467. doi: 10.1177/154193120004402123.

Baddeley, A. D., and Hitch, G. (1974). "Working Memory," in *Psychology of Learning and Motivation*, ed. G. H. Bower (Academic Press), 47–89. doi: 10.1016/S0079-7421(08)60452-1.

Bakker, A. B., and Demerouti, E. (2007). The Job Demands-Resources model: State of the art. *Journal of Managerial Psychology* 22, 309–328. doi: 10.1108/02683940710733115.

Banbury, S., and Berry, D. C. (1998). Disruption of office-related tasks by speech and office noise. *British Journal of Psychology* 89, 499–517. doi: 10.1111/j.2044-8295.1998.tb02699.x.

Bankstahl, U. S., and Görtelmeyer, R. (2013). *APSA: Attention and Performance Self-Assessment - deutsche Fassung*. , ed.Leibniz Institute for Psychology Information Trier: ZPID Available at: https://doi.org/10.23668/psycharchives.438.

Becker, R., Stasch, S.-M., Schmitz-Hübsch, A., and Fuchs, S. (2021). Quantitative scoring system to assess performance in experimental environments. in *Proceedings of the 14th International Conference on Advances in Computer-Human Interactions* (Nice, France: ThinkMind), 91–96.

Beer, A. (1852). Bestimmung der Absorption des rothen Lichts in farbigen Flüssigkeiten. *Annalen der Physik und Chemie* 86, 78–88.

Benerradi, J., Maior, H. A., Marinescu, A., Clos, J., and Wilson, M. L. (2019). Mental workload using fNIRS data from HCI tasks ground truth: Performance, evaluation, or condition. *Proceedings of the Halfway to the Future Symposium*. doi: 10.1145/3363384.3363392.

Birbaumer, N., and Schmidt, R. F. (2010). *Biologische Psychologie*. 7th ed. Berlin, Heidelberg: Springer Available at: https://doi.org/10.1007/978-3-540-95938-0.

Bird, J. J., Pritchard, M., Fratini, A., Ekárt, A., and Faria, D. R. (2021). Synthetic biological signals machine-generated by GPT-2 improve the classification of EEG and EMG through data augmentation. *IEEE Robotics and Automation Letters* 6, 3498–3504. doi: 10.1109/LRA.2021.3056355.

Bowling, N. A., Alarcon, G. M., Bragg, C. B., and Hartman, M. J. (2015). A meta-analytic examination of the potential correlates and consequences of workload. *Work & Stress* 29, 95–113. doi: 10.1080/02678373.2015.1033037.

Bradley, M. M., Miccoli, L., Escrig, M. A., and Lang, P. J. (2008). The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology* 45, 602–607. doi: 10.1111/j.1469-8986.2008.00654.x.

Brouwer, A.-M., Zander, T. O., van Erp, J. B. F., Korteling, J. E., and Bronkhorst, A. W. (2015). Using neurophysiological signals that reflect cognitive or affective state: Six recommendations to avoid common pitfalls. *Frontiers in Neuroscience* 9, 136. doi: 10.3389/fnins.2015.00136.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., and Weiss, B. (2005). A database of German emotional speech. in *Proceedings of the 9th European Conference on Speech Communication and Technology* (Lisbon, Portugal), 1520. doi: 10.21437/Interspeech.2005-446.

Bzdok, D., and Meyer-Lindenberg, A. (2018). Machine learning for precision psychiatry: Opportunities and challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* 3, 223–230. doi: 10.1016/j.bpsc.2017.11.007.

Cardoso, B., Romão, T., and Correia, N. (2013). CAAT: A discrete approach to emotion assessment. *Extended Abstracts on Human Factors in Computing Systems*, 1047–1052. doi: 10.1145/2468356.2468543.

Causse, M., Chua, Z., Peysakhovich, V., Del Campo, N., and Matton, N. (2017). Mental workload and neural efficiency quantified in the prefrontal cortex using fNIRS. *Scientific Reports* 7, 5222. doi: 10.1038/s41598-017-05378-x.

Cearns, M., Hahn, T., and Baune, B. T. (2019). Recommendations and future directions for supervised machine learning in psychiatry. *Translational Psychiatry* 9, 271. doi: 10.1038/s41398-019-0607-2.

Chakraborty, S., Aich, S., Joo, M., Sain, M., and Kim, H.-C. (2019). 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, 5397814. doi: 10.1155/2019/5397814.

Charles, R. L., and Nixon, J. (2019). Measuring mental workload using physiological measures: A systematic review. *Applied Ergonomics* 74, 221–232. doi: 10.1016/j.apergo.2018.08.028.

Chen, F., Zhou, J., Wang, Y., Yu, K., Arshad, S. Z., Khawaji, A., et al. (2016). *Robust multimodal cognitive load measurement*. Cham: Springer International Publishing doi: 10.1007/978-3-319-31700-7.

Cumming, G., and Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist* 60, 170–180. doi: 10.1037/0003-066X.60.2.170.

Curtin, A., and Ayaz, H. (2018). The age of neuroergonomics: Towards ubiquitous and continuous measurement of brain function with fNIRS. *Japanese Psychological Research* 60, 374–386. doi: 10.1111/jpr.12227.

Curtis, C. E. (2006). Prefrontal and parietal contributions to spatial working memory. *Neuroscience* 139, 173–180. doi: 10.1016/j.neuroscience.2005.04.070.

Dan, A., and Reiner, M. (2017). Real time EEG based measurements of cognitive load indicates mental states during learning. *Journal of Educational Data Mining* 9, 31–44. doi: 10.5281/zenodo.3554719.

D'Andrea-Penna, G. M., Frank, S. M., Heatherton, T. F., and Tse, P. U. (2017). Distracting tracking: Interactions between negative emotion and attentional load in multiple-object tracking. *Emotion* 17, 900–904. doi: 10.1037/emo0000329.

de Cheveigné, A., Di Liberto, G. M., Arzounian, D., Wong, D. D. E., Hjortkjær, J., Fuglsang, S., et al. (2019). Multiway canonical correlation analysis of brain data. *NeuroImage* 186, 728–740. doi: 10.1016/j.neuroimage.2018.11.026.

De Rivecourt, M., Kuperus, M. N., Post, W. J., and Mulder, L. J. M. (2008). Cardiovascular and eye activity measures as indices for momentary changes in mental effort during simulated flight. *Ergonomics* 51, 1295–1319. doi: 10.1080/00140130802120267.

Debie, E., Rojas, R. F., Fidock, J., Barlow, M., Kasmarik, K., Anavatti, S., et al. (2021). Multimodal fusion for objective assessment of cognitive workload: A review. *IEEE Transactions on Cybernetics* 51, 1542–1555. doi: 10.1109/TCYB.2019.2939399.

Dehais, F., Lafont, A., Roy, R., and Fairclough, S. (2020). A neuroergonomics approach to mental workload, engagement and human performance. *Frontiers in Neuroscience* 14, 268. doi: 10.3389/fnins.2020.00268.

Dink, J. W., and Ferguson, B. (2015). eyetrackingR: An R library for eye-tracking data analysis. Available at: http://www.eyetrackingr.com.

Dirican, A. C., and Göktürk, M. (2011). Psychophysiological measures of human cognitive states applied in human computer interaction. *Procedia Computer Science* 3, 1361–1367. doi:

10.1016/j.procs.2011.01.016.

Dolcos, F., Iordan, A. D., and Dolcos, S. (2011). Neural correlates of emotion–cognition interactions: A review of evidence from brain imaging investigations. *Journal of Cognitive Psychology* 23, 669–694. doi: 10.1080/20445911.2011.594433.

Durantin, G., Gagnon, J.-F., Tremblay, S., and Dehais, F. (2014). Using near infrared spectroscopy and heart rate variability to detect mental overload. *Behavioural Brain Research* 259, 16–23. doi: 10.1016/j.bbr.2013.10.042.

Dwyer, D. B., Falkai, P., and Koutsouleris, N. (2018). Machine learning approaches for clinical psychology and psychiatry. *Annual Review of Clinical Psychology* 14, 91–118. doi: 10.1146/annurev-clinpsy-032816-045037.

Fishburn, F. A., Ludlum, R. S., Vaidya, C. J., and Medvedev, A. V. (2019). Temporal Derivative Distribution Repair (TDDR): A motion correction method for fNIRS. *NeuroImage* 184, 171–179. doi: 10.1016/j.neuroimage.2018.09.025.

Forbes, S. (2020). PupillometryR: An R package for preparing and analysing pupillometry data. *Journal of Open Source Software* 5, 2285. doi: 10.21105/joss.02285.

Friedman, N., Fekete, T., Gal, K., and Shriki, O. (2019). EEG-based prediction of cognitive load in intelligence tests. *Frontiers in Human Neuroscience* 13, 191. doi: 10.3389/fnhum.2019.00191.

García-Pacios, J., Garcés, P., del Río, D., and Maestú, F. (2017). Tracking the effect of emotional distraction in working memory brain networks: Evidence from an MEG study. *Psychophysiology* 54, 1726–1740. doi: 10.1111/psyp.12912.

Gevins, A., and Smith, M. E. (2003). Neurophysiological measures of cognitive workload during human-computer interaction. *Theoretical Issues in Ergonomics Science* 4, 113–131. doi: 10.1080/14639220210159717.

Gottemukkula, V., and Derakhshani, R. (2011). Classification-guided feature selection for NIRS-based BCI. in *Proceedings of the 5th International IEEE/EMBS Conference on Neural Engineering 2011*, 72–75. doi: 10.1109/NER.2011.5910491.

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., et al. (2014). MNE Software for Processing MEG and EEG Data. *NeuroImage* 86, 446–460. doi: 10.1016/j.neuroimage.2013.10.027.

Hancock, P. A., and Desmond, P. A. (2001). *Stress, Workload, and Fatigue*. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Hancock, P. A., and Meshkati, N. (1988). *Human Mental Workload*. Amsterdam: North-Holland.

Hart, S. G., and Staveland, L. E. (1988). “Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research,” in *Advances in Psychology*, eds. P. A. Hancock and N. Meshkati (North-Holland), 139–183. doi: 10.1016/S0166-4115(08)62386-9.

Hartmann, A. S., Rief, W., and Hilbert, A. (2011). Psychometric properties of the German version of

the Barratt Impulsiveness Scale, Version 11 (BIS–11) for adolescents. *Perceptual and Motor Skills* 112, 353–368. doi: 10.2466/08.09.10.PMS.112.2.353-368.

Herff, C., Heger, D., Fortmann, O., Hennrich, J., Putze, F., and Schultz, T. (2014). 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.

Herms, R., Wirzberger, M., Eibl, M., and Rey, G. D. (2018). CoLoSS: Cognitive load corpus with speech and performance data from a symbol-digit dual-task. in *Proceedings of the 11th International Conference on Language Resources and Evaluation* (Miyazaki, Japan: European Language Resources Association). Available at: https://aclanthology.org/L18-1681.

Hoc, J.-M. (2001). Towards ecological validity of research in cognitive ergonomics. *Theoretical Issues in Ergonomics Science* 2, 278–288. doi: 10.1080/14639220110104970.

Hosseini, S. M. H., Bruno, J. L., Baker, J. M., Gundran, A., Harbott, L. K., Gerdes, J. C., et al. (2017). Neural, physiological, and behavioral correlates of visuomotor cognitive load. *Scientific Reports* 7, 8866. doi: 10.1038/s41598-017-07897-z.

Huppert, T. J., Diamond, S. G., Franceschini, M. A., and Boas, D. A. (2009). HomER: A review of time-series analysis methods for near-infrared spectroscopy of the brain. *Applied Optics* 48, 280–298. doi: 10.1364/AO.48.00D280.

Izzetoglu, K., Bunce, S., Izzetoglu, M., Onaral, B., and Pourrezaei, K. (2003). fNIR spectroscopy as a measure of cognitive task load. in *Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 3431-3434 Vol.4. doi: 10.1109/IEMBS.2003.1280883.

Jackson, I., and Sirois, S. (2009). Infant cognition: Going full factorial with pupil dilation. *Developmental Ccience* 12, 670–679. doi: 10.1111/j.1467-7687.2008.00805.x.

Keles, H. O., Cengiz, C., Demiral, I., Ozmen, M. M., and Omurtag, A. (2021). High density optical neuroimaging predicts surgeons's subjective experience and skill levels. *PLOS ONE* 16, e0247117. doi: 10.1371/journal.pone.0247117.

Klimesch, W. (2011). Evoked alpha and early access to the knowledge system: The P1 inhibition timing hypothesis. *Brain Research* 1408, 52–71. doi: 10.1016/j.brainres.2011.06.003.

Kramer, A. F. (1991). "Physiological metrics of mental workload: A review of recent progress," in *Multiple-task performance*, ed. D. L. Damos (London: CRC Press), 279–328. doi: 10.1201/9781003069447.

Ladouce, S., Donaldson, D. I., Dudchenko, P. A., and Ietswaart, M. (2017). Understanding minds in real-world environments: Toward a mobile cognition approach. *Frontiers in Human Neuroscience* 10, 694. doi: 10.3389/fnhum.2016.00694.

Lashgari, E., Liang, D., and Maoz, U. (2020). Data augmentation for deep-learning-based electroencephalography. *Journal of Neuroscience Methods* 346, 108885. doi: 10.1016/j.jneumeth.2020.108885.

Laux, L., Glanzmann, P., Schaffner, P., and Spielberger, C. D. (1981). *Das State-Trait-Angstinventar*. Weinheim: Beltz.

Lavie, N. (2010). Attention, distraction, and cognitive control under load. *Current Directions in Psychological Science* 19, 143–148. doi: 10.1177/0963721410370295.

Lawhern, V. J., Solon, A. J., Waytowich, N. R., Gordon, S. M., Hung, C. P., and Lance, B. J. (2018). EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering* 15, 056013. doi: 10.1088/1741-2552/aace8c.

Lemm, S., Blankertz, B., Dickhaus, T., and Müller, K.-R. (2011). Introduction to machine learning for brain imaging. *NeuroImage* 56, 387–399. doi: 10.1016/j.neuroimage.2010.11.004.

Liebl, A., Haller, J., Jödicke, B., Baumgartner, H., Schlittmeier, S., and Hellbrück, J. (2012). Combined effects of acoustic and visual distraction on cognitive performance and well-being. *Applied Ergonomics* 43, 424–434. doi: 10.1016/j.apergo.2011.06.017.

Liu, Y., Lan, Z., Cui, J., Sourina, O., and Müller-Wittig, W. (2019). EEG-based cross-subject mental fatigue recognition. in *Proceedings of the International Conference on Cyberworlds 2019*, 247–252. doi: 10.1109/CW.2019.00048.

Lotte, F., Bougrain, L., Cichocki, A., Clerc, M., Congedo, M., Rakotomamonjy, A., et al. (2018). A review of classification algorithms for EEG-based brain–computer interfaces: A 10 year update. *Journal of Neural Engineering* 15, 031005. doi: 10.1088/1741-2552/aab2f2.

Luke, R., Larson, E. D., Shader, M. J., Innes-Brown, H., Van Yper, L., Lee, A. K. C., et al. (2021). Analysis methods for measuring passive auditory fNIRS responses generated by a block-design paradigm. *Neurophotonics* 8, 1–18. doi: 10.1117/1.NPh.8.2.025008.

Lyu, B., Pham, T., Blaney, G., Haga, Z., Sassaroli, A., Fantini, S., et al. (2021). Domain adaptation for robust workload level alignment between sessions and subjects using fNIRS. *Journal of Biomedical Optics* 26, 1–21. doi: 10.1117/1.JBO.26.2.022908.

Marquart, G., Cabrall, C., and de Winter, J. (2015). Review of eye-related measures of drivers' mental workload. *Procedia Manufacturing* 3, 2854–2861. doi: 10.1016/j.promfg.2015.07.783.

Martínez-Vázquez, P., and Gail, A. (2018). Directed interaction between monkey premotor and posterior parietal cortex during motor-goal retrieval from working memory. *Cerebral Cortex* 28, 1866–1881. doi: 10.1093/cercor/bhy035.

Matthews, G., De Winter, J., and Hancock, P. A. (2020). What do subjective workload scales really measure? Operational and representational solutions to divergence of workload measures. *Theoretical Issues in Ergonomics Science* 21, 369–396. doi: 10.1080/1463922X.2018.1547459.

Matthews, R., McDonald, N. J., and Trejo, L. J. (2005). "Psycho-physiological sensor techniques: An overview," in *Foundations of Augmented Cognition*, ed. D. D. Schmorrow (CRC Press), 263–272. doi: 10.1201/9781482289701.

Midha, S., Maior, H. A., Wilson, M. L., and Sharples, S. (2021). Measuring mental workload variations in office work tasks using fNIRS. *International Journal of Human-Computer Studies* 147,

1036 102580. doi: 10.1016/j.ijhcs.2020.102580.

1037 Miller, E. K., Freedman, D. J., and Wallis, J. D. (2002). The prefrontal cortex: Categories, concepts
1038 and cognition. *Philosophical Transactions of the Royal Society of London* 357, 1123–1136. doi:
1039 10.1098/rstb.2002.1099.

1040 Minkley, N., Xu, K. M., and Krell, M. (2021). Analyzing relationships between causal and
1041 assessment factors of cognitive load: Associations between objective and subjective measures of
1042 cognitive load, stress, interest, and self-concept. *Frontiers in Education* 6. Available at:
1043 https://www.frontiersin.org/article/10.3389/feduc.2021.632907.

1044 Neisser, U. (1976). *Cognition and Reality: Principles and Implications of Cognitive Psychology*. San
1045 Francisco, CA, USA: W. H. Freeman.

1046 Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular
1047 reference to demand characteristics and their implications. *American Psychologist* 17, 776–783. doi:
1048 10.1037/h0043424.

1049 Orrù, G., Monaro, M., Conversano, C., Gemignani, A., and Sartori, G. (2020). Machine learning in
1050 psychometrics and psychological research. *Frontiers in Psychology* 10, 2970. doi:
1051 10.3389/fpsyg.2019.02970.

1052 Paas, F., Tuovinen, J. E., Tabbers, H., and Van Gerven, P. W. M. (2003). Cognitive load
1053 measurement as a means to advance cognitive load theory. *null* 38, 63–71. doi:
1054 10.1207/S15326985EP3801_8.

1055 Pacific Science & Engineering Group (2003). *Warship Commander 4.4*. San Diego, CA, USA.

1056 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-
1057 learn: Machine Learning in Python. *Journal of Machine Learning Research* 12, 2825–2830.

1058 Pollonini, L., Olds, C., Abaya, H., Bortfeld, H., Beauchamp, M. S., and Oghalai, J. S. (2014).
1059 Auditory cortex activation to natural speech and simulated cochlear implant speech measured with
1060 functional near-infrared spectroscopy. *Hearing Research* 309, 84–93. doi:
1061 10.1016/j.heares.2013.11.007.

1062 Rammstedt, B., and John, O. P. (2005). Kurzversion des Big Five Inventory (BFI-K): *Diagnostica*
1063 51, 195–206. doi: 10.1026/0012-1924.51.4.195.

1064 Ranchet, M., Morgan, J. C., Akinwuntan, A. E., and Devos, H. (2017). Cognitive workload across
1065 the spectrum of cognitive impairments: A systematic review of physiological measures.
1066 *Neuroscience & Biobehavioral Reviews* 80, 516–537. doi: 10.1016/j.neubiorev.2017.07.001.

1067 Raschka, S. (2018). MLxtend: Providing machine learning and data science utilities and extensions to
1068 Python's scientific computing stack. *Journal of Open Source Software* 3, 638. doi:
1069 10.21105/joss.00638.

1070 Romine, W. L., Schroeder, N. L., Graft, J., Yang, F., Sadeghi, R., Zabihimayvan, M., et al. (2020).
1071 Using machine learning to train a wearable device for measuring students' cognitive load during
1072 problem-solving activities based on electrodermal activity, body temperature, and heart rate:

Development of a cognitive load tracker for both personal and classroom use. *Sensors* 20. doi: 10.3390/s20174833.

Saager, R. B., and Berger, A. J. (2005). Direct characterization and removal of interfering absorption trends in two-layer turbid media. *Journal of the Optical Society of America A* 22, 1874–1882. doi: 10.1364/JOSAA.22.001874.

Scheunemann, J., Unni, A., Ihme, K., Jipp, M., and Rieger, J. W. (2019). Demonstrating brain-level interactions between visuospatial attentional demands and working memory load while driving using functional near-infrared spectroscopy. *Frontiers in Human Neuroscience* 12, 542. doi: 10.3389/fnhum.2018.00542.

Schiratti, J.-B., Le Douget, J.-E., Le Van Quyen, M., Essid, S., and Gramfort, A. (2018). An ensemble learning approach to detect epileptic seizures from long intracranial EEG recordings. in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing 2018*, 856–860. doi: 10.1109/ICASSP.2018.8461489.

Schneider, S., Wirzberger, M., and Rey, G. D. (2019). The moderating role of arousal on the seductive detail effect in a multimedia learning setting. *Applied Cognitive Psychology* 33, 71–84. doi: 10.1002/acp.3473.

Schweizer, S., Satpute, A. B., Atzil, S., Field, A. P., Hitchcock, C., Black, M., et al. (2019). The impact of affective information on working memory: A pair of meta-analytic reviews of behavioral and neuroimaging evidence. *Psychological Bulletin* 145, 566–609. doi: 10.1037/bul0000193.

Sörqvist, P., Dahlström, Ö., Karlsson, T., and Rönnberg, J. (2016). Concentration: The neural underpinnings of how cognitive load shields against distraction. *Frontiers in Human Neuroscience* 10, 221. doi: 10.3389/fnhum.2016.00221.

St John, M., Kobus, D. A., and Morrison, J. G. (2003). DARPA augmented cognition technical integration experiment (TIE). San Diego, CA, USA: Pacific Science and Engineering Group.

Taelman, J., Vandeput, S., Spaepen, A., and Van Huffel, S. (2009). Influence of mental stress on heart rate and heart rate variability. in *4th European Conference of the International Federation for Medical and Biological Engineering*, eds. J. Vander Sloten, P. Verdonck, M. Nyssen, and J. Haueisen (Berlin, Heidelberg: Springer), 1366–1369.

Tao, D., Tan, H., Wang, H., Zhang, X., Qu, X., and Zhang, T. (2019). A systematic review of physiological measures of mental workload. *International Journal of Environmental Research and Public Health* 16, 2716. doi: 10.3390/ijerph16152716.

Toet, A., Kaneko, D., Ushiama, S., Hoving, S., de Kruijf, I., Brouwer, A.-M., et al. (2018). EmojiGrid: A 2D pictorial scale for the assessment of food elicited emotions. *Frontiers in Psychology* 9, 2396. doi: 10.3389/fpsyg.2018.02396.

Uludağ, K., and Roebroeck, A. (2014). General overview on the merits of multimodal neuroimaging data fusion. *NeuroImage* 102, 3–10. doi: 10.1016/j.neuroimage.2014.05.018.

Unni, A., Ihme, K., Jipp, M., and Rieger, J. W. (2017). 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* 11, 167. doi: 10.3389/fnhum.2017.00167.

Vanneste, P., Raes, A., Morton, J., Bombeke, K., Van Acker, B. B., Larmuseau, C., et al. (2021). Towards measuring cognitive load through multimodal physiological data. *Cognition, Technology & Work* 23, 567–585. doi: 10.1007/s10111-020-00641-0.

von der Malsburg, T. (2015). saccades: Detection of fixations in eye-tracking data. Available at: https://github.com/tmalsburg/saccades.

von Lühmann, A. (2018). Multimodal instrumentation and methods for neurotechnology out of the lab. *Fakultät IV - Elektrotechnik und Informatik*. doi: 10.14279/depositonce-7445.

von Lühmann, A., Boukouvalas, Z., Müller, K.-R., and Adalı, T. (2019). A new blind source separation framework for signal analysis and artifact rejection in functional near-infrared spectroscopy. *NeuroImage* 200, 72–88. doi: 10.1016/j.neuroimage.2019.06.021.

Vu, M.-A. T., Adalı, T., Ba, D., Buzsáki, G., Carlson, D., Heller, K., et al. (2018). A shared vision for machine learning in neuroscience. *The Journal of Neuroscience* 38, 1601. doi: 10.1523/JNEUROSCI.0508-17.2018.

Vuilleumier, P., and Schwartz, S. (2001). Emotional facial expressions capture attention. *Neurology* 56, 153–158. doi: 10.1212/WNL.56.2.153.

Waytowich, N. R., Lawhern, V. J., Bohannon, A. W., Ball, K. R., and Lance, B. J. (2016). Spectral transfer learning using information geometry for a user-independent brain-computer interface. *Frontiers in Neuroscience* 10, 430. doi: 10.3389/fnins.2016.00430.

Wierwille, W. W. (1979). Physiological measures of aircrew mental workload. *Human Factors* 21, 575–593. doi: 10.1177/001872087902100504.

Wirzberger, M., Herms, R., Esmaeili Bijarsari, S., Eibl, M., and Rey, G. D. (2018). Schema-related cognitive load influences performance, speech, and physiology in a dual-task setting: A continuous multi-measure approach. *Cognitive Research: Principles and Implications* 3, 46. doi: 10.1186/s41235-018-0138-z.

Yap, T. F., Epps, J., Ambikairajah, E., and Choi, E. H. C. (2015). Voice source under cognitive load: Effects and classification. *Speech Communication* 72, 74–95. doi: 10.1016/j.specom.2015.05.007.

Young, M. S., Brookhuis, K. A., Wickens, C. D., and Hancock, P. A. (2015). State of science: Mental workload in ergonomics. *Ergonomics* 58, 1–17. doi: 10.1080/00140139.2014.956151.

Yücel, M. A., Lühmann, A. v., Scholkmann, F., Gervain, J., Dan, I., Ayaz, H., et al. (2021). Best practices for fNIRS publications. *Neurophotonics* 8, 1–34. doi: 10.1117/1.NPh.8.1.012101.

Zanini, R. A., and Colombini, E. L. (2020). Parkinson's disease EMG data augmentation and simulation with DCGANs and Style Transfer. *Sensors* 20, 2605. doi: 10.3390/s20092605.

Zhang, Y.-D., Dong, Z., Wang, S.-H., Yu, X., Yao, X., Zhou, Q., et al. (2020). Advances in multimodal data fusion in neuroimaging: Overview, challenges, and novel orientation. *Information Fusion* 64, 149–187. doi: 10.1016/j.inffus.2020.07.006.

Zheng, R. Z. (2017). *Cognitive load measurement and application: A theoretical framework for meaningful research and practice*. New York, NY, US: Routledge Available at: https://doi.org/10.4324/9781315296258.

Zimeo Morais, G. A., Balardin, J. B., and Sato, J. R. (2018). fNIRS Optodes' Location Decider (fOLD): A toolbox for probe arrangement guided by brain regions-of-interest. *Scientific Reports* 8, 3341. doi: 10.1038/s41598-018-21716-z.

**Table 1**

***Included Features per Modality***

| Modality | Features |
|---|---|
| **Brain Activity** | Mean, standard deviation, peak-to-peak (PTP) amplitude, skewness, and kurtosis of the 81 optical channels |
| **Physiology** | |
| Heart Rate | Mean, standard deviation, skewness, and kurtosis of heart rate<br>Mean, standard deviation, skewness, and kurtosis of heart rate variability |
| Respiration | Mean, standard deviation, skewness, and kurtosis of respiration rate<br>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 |

**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, participants had to warn that drone; when an unregistered drone entered the red area, participants had to repel it. Right side of the screen: Participants had to request codes and pictures of unknown flying objects and then classify them as birds, registered drones, or unregistered drones.

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

**Figure 3.** fNIRS optodes' location. Montage of optodes on fNIRS cap on a standard 10-20 EEG system, red optodes: near-infrared light-emitting sources, blue optodes: detectors, green lines: long channels, blue lines: short channels. The setup resulted in 41 (source-detector-pairs) × 2 (wavelengths) = 82 optical channels of interest.

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

**Figure 5.** Prediction of the subjectively perceived mental effort ($F_1$ scores in training and test set) based on a median split; validation set: $N = 1$. Bootstrapped 95% confidence intervals (CI; 5000 iterations) of the $F_1$ scores for the training set (left, orange) and the test set (right, blue) of the different unimodal and multimodal prediction 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 solid black line represents the mean, whereas the dashed grey line in the box represents the median. The continuous grey dashed line shows the upper boundary of the CI of the dummy classifier and is hence an indicator of whether the model performance is significantly better than the dummy classifier.

**Figure 6.** Weights and procedure of a unimodal voting classifier to predict subjectively perceived mental effort based on a median split; validation set: $N = 1$. The weights and the procedure were selected by the means of a cross-validated randomized grid search with 100 iterations. Possible weights were either 0, 1, or 2.

**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 Classifiers" we selected the classifier (LDA, LR, SVM, KNN, RFC, or GNB) with the best $F_1$ score for each subject.

**Figure 8.** Weights and procedure of a multimodal voting classifier to predict subjectively perceived mental effort based; validation set: $N = 1$. The weights and the procedure were selected by the means of a cross-validated randomized grid search with 100 iterations. Possible weights were either 0, 1, or 2. **(B)** shows the allocation of weights when performance measures are not included in the multimodal classification.

## Data Availability Statement

The datasets analyzed for this study as well as the code can be found in a publicly accessible OSF repository (https://osf.io/9dbcj/?view_only=9d0b718ee8984f568f09ca5b7db3c6fb).

## Ethics Statement

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

## Conflict 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.

## Author Contributions

KL designed the study and contributed to the data acquisition. SG performed the analyses and visualizations and wrote the original draft manuscript. SG, KL, MW, and MV performed iterative reviews and edits. KL and MV supervised SG. All authors contributed to the article and approved the submitted version.

## 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«).

## 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. We would like to thank Kristina Kögler and Kathrin Pape for the fruitful discussions concerning the experimental design and emotion induction in naturalistic scenarios.