Emotions Matter: Towards Personalizing Human-System Interactions Using a Two-layer Multimodal Approach

Apostolos Kalatzis
Montana State University
Bozeman, Montana
apostoloskalatzis@smontana.edu

Vishnunarayan Girishan Prabhun
Clemson University
Clemson, South Carolina
vgirish@clemson.edu

Saidur Rahman
Montana State University
Bozeman, Montana
saidurrahman@montana.edu

Laura Stanley
Montana State University
Bozeman, Montana
Laura.stanley@montana.edu

Mike Wittie
Montana State University
Bozeman, Montana
mike.wittie@montana.edu

ABSTRACT

Monitoring and predicting user task performance is critical as it provides valuable insights for developing personalized human-system interactions. Key factors that impact task performance include cognitive workload, physiological responses, and affective states. However, a lack of consideration of any of these factors could lead to inaccurate task performance prediction because of their interplay. To address this challenge, we developed a novel hierarchical machine learning approach that considers these three factors to predict task performance. We exposed twenty-eight participants to a two-step experimental study. The first step aimed to induce different affective states using a validated video database. The second step required participants to perform validated low and high cognitive workload-inducing tasks. To evaluate the performance, we compared the models developed using the hierarchical approach that uses emotional and physiological information, to models that use only physiological information. We observed that our proposed approach always outperformed the models that only use physiological information to predict task performance by achieving a better average person independent mean absolute error. However, information gained across various models using the hierarchical approach was not linear. Additionally, we found that the top predictors for each model varied, and the model with the highest information gain included emotional features. These findings suggest the importance of choosing the appropriate machine learning model and predictors for building robust models for predicting task performance.

KEYWORDS

Machine learning; Affective Computing; Task performance; Physiological signals; Cognitive workload

1 INTRODUCTION

Humans perform various tasks on a day-to-day basis, and each task requires a level of cognitive resources. The cognitive resources required to complete a task vary based on the task complexity and other factors. The ratio of resources needed to perform a task at hand to the total resources the human has available to dedicate to the task is defined as cognitive workload. Cognitive workload is as a significant factor influencing human performance [4, 24, 33, 40, 56]. Hence, investigating and understanding factors affecting cognitive workload can provide significant insights to develop personalized human-machine interactions and collaborations including driving systems [39], nuclear power plants [25], and flight stations [3] that enhance the performance. With the rapid growth of technology, the relevance of cognitive workload and monitoring task performance has increased significantly across multiple occupational areas. Irrespective of the area the consensus is that either high or excessively low levels of cognitive workload negatively influence operators' work performance [31, 52].

According to Wickens' multiple resource theory, humans have only a limited amount of cognitive resources to dedicate to a task, and when the task at hand exceeds the available resources it leads to inefficiency and deteriorated performance [54]. On the other hand, when a task requires less resources it can distract users from the primary task leading to a lack of vigilance and result in a subpar performance [56]. Monitoring cognitive workload becomes particularly important to assess safety-critical systems, where suboptimal performance could result in errors, potential accidents, and worker dissatisfaction [52].

However, monitoring cognitive workload is not simple as it is a multidimensional phenomenon influenced by various factors and demands a comprehensive understanding. Abbass et al. created an equation to define cognitive load where the amount of mental resources needed to perform the task is calculated by adding the workload related to the work environment (work load) and that imposed by exogenous environmental factors (environmental load) [1]. Although this equation considers some personal life experiences, it does not account for the user's emotions in the equation. Personal life experiences, including emotions, significantly impact cognitive workload and performance as reported in prior studies [2, 19, 50]. Emotions can be defined as functional behaviors influenced by thoughts, stimuli, and other factors that induce neurophysiological changes in the human body. Because of this interconnectedness, emotions can interfere with cognitive resources and can lead to poor task performance [2, 19].

Emotions, also known as affective states, can be quantified using a validated model based on valence and arousal [41]. Valence represents the affective state, which could be positive (pleasant) or negative (unpleasant) in response to a particular event or situation. Whereas arousal measures the affective state's intensity, which can vary between highly calm to highly excited or alert. Although valence and arousal are subjective metrics, they are associated with physiological measures that objectively represent users' emotions [29, 51]. Studies have independently investigated the relationship between cognitive workload and task performance [30] and the relationship between emotions and cognitive workload [5, 43]. However, to our knowledge, none of the studies have investigated the interplay between emotions, physiological measures, and cognitive workload for predicting task performance.

We hypothesize that the interplay of these three factors directly impacts task performance and can provide better insight for predicting task performance. To evaluate this hypothesis, we propose a novel two-layer machine learning approach that leverages the interplay between physiological responses, emotional state, and cognitive workload to predict user task performance. In the machine learning model, the first layer uses an emotion assessment model that predicts the valence and arousal of the user while the second layer predicts the task performance taking into consideration physiological responses along with the predicted level of valence and arousal from the first layer. The contribution of this research is twofold: (a) first, we investigate the interplay of physiological responses, emotional state, and cognitive workload in predicting a user's task performance (b) second, we develop a reliable method and robust model for predicting a user's task performance which could significantly impact the use of future technologies, specifically those which require personalized human-machine interactions and collaboration such as smart manufacturing.

2 BACKGROUND AND RELATED WORK

2.1 Association between Emotions and Cognitive Workload

The literature identifies the connection between emotions and cognitive workload where emotions affect the cognitive processing efficiency. There are several ways to consider the effects of emotions on cognitive load and the most common approach considers emotions as a source of extraneous cognitive load [19]. In this approach, triggered emotions create demands on cognitive resources that need to be processed. These demands are not relevant to the task goals and processing these can lead to extraneous cognitive load. Additionally, both positive or negative emotions can cause extra task processing and result in increased cognitive load [2, 19, 50]. Further, the relation between emotions and cognitive workload is linked to intrinsic cognitive load. This assumes emotional regulation as part of the task process and consider emotions as a source of intrinsic cognitive load [52]. Another way to consider the effect of emotions on cognitive load is by measuring the effect of emotion on motivation. Studies have observed that positive or negative emotions can trigger intrinsic motivation and result in an increased cognitive demand [34].

2.2 Heart Rate and Respiration Rate as Cognitive Workload and Emotion measures

Cognitive workload and emotions have a significant effect on human performance, however, quantifying these objectively is difficult [31, 52]. Various studies have demonstrated that physiological measures can act as a surrogate of cognitive workload and affective states [12, 14, 26]. The physiological measures that capture the change in cognitive workload, and emotions can be represented as a function of the autonomic nervous system (ANS). The ANS consists of two major systems: the sympathetic and parasympathetic systems [13, 47]. The sympathetic system is our "fight and flight" response, which activates during the stress, and exertion states. In contrast, the parasympathetic system is our "rest and digest" response, which is dominant during a relaxed state [11, 46]. The most commonly assessed indices of ANS are based on cardiovascular and respiration activity [7, 8, 21]. Prior studies have observed cardiovascular activity and respiration rate are critical indices that can explain the balance or imbalance in ANS as a result of cognitive workload [6, 17, 22]. Task type, task load, and task difficulty can be recognized using physiological features by observing differences in cognitive state. Researchers have reported a positive correlation between heart rate (HR) levels, cognitive load, and task difficulty [6, 17, 22]. Additionally, time domain and frequency domain metrics of heart rate variability (HRV) have been investigated to measure the cognitive workload. Studies have consistently observed a decrease in the time domain measures(i.e., mean HRV, mean NN) and an increase in the frequency domain metric (i.e., LF/HF ratio) while performing tasks that require higher mental demand. These observations suggest an increase in sympathetic activity [16, 18, 32, 55]. Similarly, respiration rate (RSP) represents the rate at which a person breathes per minute and is a crucial determinant of cognitive workload where RSP increases as the complexity of tasks increases [20, 23].

Moreover, emotions are associated with a change in neurophysiologic activities, and prior studies have noted a distinct effect on human physiology. HRV and respiration have been widely used as indicators of physiological internal states associated with emotion or affect [27, 28, 36, 44]. In recent years, several studies explored the feasibility of detecting emotions using HRV features. In the work of Valenza et al. (2014), linear and nonlinear HRV computed to create a real-time emotion recognition system able to identify two levels of arousal and valence (low-medium and medium-high) [51]. Guo et al. (2016) monitored two emotion states (positive/negative) using time-domain, frequency-domain, Poincare, and statistical HRV features [29]. In the work of Cheng et al. (2017) a real-time negative emotion detection method introduced using linear-derived features, nonlinear-derived features, time-domain features, and time-frequency domain features [15].

2.3 Machine Learning Detection of Task Performance

In recent years advances in machine learning have piqued the interest of researchers in using algorithms to classify cognitive workload based on physiological responses [15, 45]. Although there is much work focusing on predicting the cognitive workload from physiological signals, only a few studies aim to directly predict a user's performance on a specific task. Papakostas et al, considered physiological data to train machine learning models to predict the user's task performance in a sequence learning task. This study utilized data collected from 69 participants during a working memory task to evaluate the ability of a human to remember and repeat a sequence of items (e.g., letters, numbers, actions). The classifier algorithm evaluated using a 10-fold cross-validation, and the authors achieved an accuracy of 74% [43].

Another recent study aimed to predict the user's task performance using a multimodal approach during a sequence learning task where the participants had to remember and repeat a sequence of characters [5]. The authors created a scoring scheme to predict the success or failure of the task based on three different modalities; emotions based on facial expressions, emotions based on posture, and task performance based on physiological data. The authors achieved an accuracy of 87.5% from the combined modalities using a neural network. In both studies, authors predicted the task performance of each trial as a success or fail (binary). Papakostas et al., considered only physiological measures whereas Babu et al., considered both physiological measures and emotional state. However, this study uses emotions based on facial recognition. The binary outcome of these studies for task performance limits the application of these models for addressing real-life scenarios as task performance is usually quantified on a continuous scale [53, 54]. Our research aims to address this gap by building a hierarchical model-based on user's emotions and physiological activity to predict task performance on a continuous scale.

3 METHODS

3.1 Study

In this study, the participants followed a two-step experimental procedure as seen in the figure 1. The first step aimed to induce different levels of arousal using validated affective stimuli [35]. In the next step, the same participants were exposed to two validated tasks designed to induce low and high cognitive workloads [42]. The order of the steps was counterbalanced to control for order effects, and upon completing each step and there was a half-hour break.

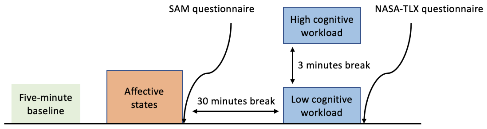


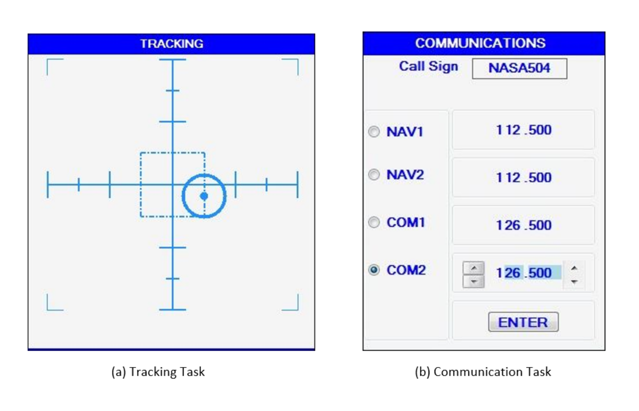
Figure 1: Study Protocol

3.2 Participants

Participants for this study included 28 students (13M, 15F, Mean = 22.92 ± 5.09) attending a public university. As a result of the university policies associated with COVID-19 restrictions, participants were recruited by convenient sampling, and participation in the study was voluntary. All participants included in the study signed a consent form and were subject to the experiment protocols approved by university's Institutional Review Board under IRBNO: AK030220-EX. There were no dropouts in this study and participants were compensated with a $25 Amazon gift card for their time.

3.3 Cognitive Workload Manipulation

3.3.1 Procedure. The two tasks which aimed to induce cognitive load were presented on a display with 1680 x 1050 resolution. Participants were seated in a rigid but comfortable posture with feet placed flat on the floor in a quiet room with normal light and temperature. Before exposing the participants to the cognitive workload tasks, physiological measures were collected for five minutes during which the participant remained silent and seated in a comfortable posture at the testing station. The data collected during this period represented the baseline metric and no participatory tasks were administered during this time. Following the baseline, the participants performed a practice trial to familiarize themselves with the nature of the task and eliminate the potential for participatory errors associated with a lack of confidence with the interface or the need to learn fundamental tasks extemporaneously. Once the participants were comfortable, they were asked to rest for three minutes to control for the effect of training on physiological data [52].



(a) Tracking Task

(b) Communication Task

Figure 2: Cognitive Workload Inducing Tasks

To induce the two levels of cognitive workload, the validated NASA MATB II application was utilized as seen in figure 2 [42]. The low cognitive workload consisted of a tracking (primary) task that lasted for 15 minutes. Figure 2 (a) represents the tracking task, where the participants had to use a joystick to keep a moving target (circle) in the center of an inner box presented on the desktop screen (Figure 2 (a)). The root mean square deviation between the center dot of the circle and the center of the small box represents the task performance. For the high cognitive workload task, the participants had to perform the two tasks simultaneously, a primary task and a secondary task. The primary task was the same tracking task, and for the secondary task, we utilized a communication task, as seen in Figure 2 (b), where participants listened to messages every 30 seconds (like air traffic control requests) and were required to adjust the radio to a specific frequency. The order of the tasks was counterbalanced to control for order effects and upon completing each task, the participants were given a three-minute rest period to control the carry-forward effect of one task on the physiological data. During the experience and the rest period, one of the researchers continuously monitored the physiological data to assure that they were collected and transmitted without any issues.

3.3.2 Dependent Measurements. The response variables considered include: perceived cognitive workload, heart rate, respiration rate activity measurements, and task performance accuracy. The NASA-TLX questionnaire captured the perceived cognitive cognitive workload after each task. The physiological data were collected using the clinically validated Biopac® MP160 system. The specific device is portable, non-obtrusive, and allows for modular data acquisition. For collecting the ECG signals, we used 3-lead chest electrodes, and a respiration belt that went around the chest was utilized for collecting the respiration signals. The electrocardiogram (ECG) and respiration rate (RSP) signals were collected at a sampling rate of 1000Hz. Task performance was quantified using an accuracy metric that was obtained by calculating the root mean square deviation from the target center point in pixel units.

3.4 Emotional Stimuli

3.4.1 Procedure. There are several methods in research to elicit affective states in the laboratory. A few most common and accepted stimuli to evoke emotions are pictures [37], films [49], and music videos [35]. In this study, we selected twelve videos from the publicly available dataset: *Database for Emotion Analysis using Physiological Signals (DEAP)* [35]. The emotional ratings of these videos were distributed uniformly over the Arousal-Valence plane with four music videos to correspond to each of the four quadrants of the 2D emotional model. Participants for this study were the same set of participants that were exposed to the two cognitive tasks that aimed to induce a varying workload. Using the same participant pool was critical as emotions and physiological measures are subject-dependent. Prior to exposing the participants to the affective stimuli, initial quantitative measures were collected during a five-minute period in which the participant remained silent and seated in a comfortable posture at the testing station. Following the baseline, the participants were exposed to the emotional stimuli. The twelve videos were presented in twelve trials. Each trial consisted of a one-minute display of the music video and between each trial, the participants were given a one-minute rest period to control the carry-forward effect of one experience on the physiological data.

3.4.2 Dependent Measurements. The response variables considered include: valence and arousal perceptions and heart rate and respiration rate activity measurements. The valence and arousal perceptions were captured through the self-assessment manikin (SAM) questionnaire after completing each trial. SAM is a self-assessment questionnaire that captures the user's perception of valence and arousal associated with stimuli [9]. The SAM contains a series of images for each dimension (valence, arousal) that range along a 9-point-scale. Each image defines a point in the scale for each dimension: valence ranges from 1 -'unpleasant' to 9 - 'pleasant', arousal ranges from 1 -'calm' to 9 - 'aroused'.

3.5 Feature Extraction

To extract the ECG signal features, we first detected the Q wave, R wave, S wave (QRS) complex for each window and then derived the time series metric of RR interval data [12]. Next, we performed a power spectrum analysis using wavelet transformation, a time-frequency analysis method to scale the decomposed ECG signal into different frequency band signals [14]. To derive the respiration rate metrics, we used low-pass signal filtering and the detection of zero-crossings of the filtered signal on the respiration signals. Further, to obtain the iterative breath interval (IBI), we calculated the respiration rate variability (RRV) by recording the time stamp at the peak of each breath pulse and then subtracted the subsequent peaks to obtain the breath-to-breath (BB) interval [38]. Finally, the quality, scaling, and correlation properties of the ECG and RSP signals were assessed to extract the nonlinear-domain features. In total, we extracted five HRV features in time (mean HR, SD HR, RMSSD, mean NN) [48] and frequency domain (LF/HF) [48]. Additionally, we extracted five RRV features in time domain (mean RSP, SD RSP, RMSSD, mean BB, SD BB)) [38].

3.5.1 Emotional Assessment. For the emotional assessment model development, each video trial serves as a window for feature extraction. HR and HRV features were extracted from the ECG signal and RSP and RRV features from the RSP signal for each window.

3.5.2 Task Performance. For the task performance model development, we extract physiological data using a window size of 300,000 data points, which corresponds to five minutes, with a sliding window of size 30,000 points corresponding to 30 seconds.

3.6 Machine Learning Model Development

We chose several machine learning models to have linear and nonlinear ranges and models sensitive to smaller sample sizes. For all the models we followed the same procedure to select the best hyperparameters. we defined a grid search with a range of the hyperparameters and then we created subject folds in order to identify these parameters based on the leave one subject out cross-validation technique. The same technique wa used to evaluate our models. This technique ensures that the best hyperparameters were selected to train these models with respect to the evaluation technique. Therefore, all the models were optimized to ensure the best performing results.

3.6.1 Emotional Assessment. After extracting the features, we started the machine learning analysis, in which the goal was to predict the affective state based on the change in physiological signals during the music videos. For this, we utilized three machine learning algorithms: support vector regressor (SVR), decision tree regressor (DTR), and random forest regression (RFR). To evaluate the machine learning models, we trained the model and then performed leave-one-subject-out (LOSO) cross-validation, where data from one participant is randomly selected for testing purposes while the data from other participants are used for training the model. This process repeated until all the participants were a part of the test dataset. We calculated the mean absolute error (MAE) and the root mean square error (RMSE) for each participant.

3.6.2 Task Performance Prediction. Two-layer models: As discussed in the previous section using only physiological measures for predicting task performance could lead to poor prediction results as emotion is a key factor influencing task performance. Hence, we propose a novel two-layer approach where in the first step, we use an emotion assessment model for predicting the arousal and valence level of the user during the cognitive tasks. In the next step, the predicted arousal value along with HRV and RRV features extracted from the physiological data during the cognitive tasks were combined to serve as a feature set for predicting the user's task performance. For predicting the task performance we trained three machine-learning algorithms: DTR, SVR, and multilayer perceptron regressor (MLPR). All three algorithms were validated using the LOSO cross-validation and mean absolute error (MAE) and the root mean square error (RMSE) were calculated for each participant.

Single-Layer Model: In this approach we predict the task performance considering only HR, RR, HRV, and RRV features collected during the cognitive tasks. The single-layer approach utilizes the same models as the two-layer and the same evaluation metrics.

3.7 Significant Predictors Calculation

We quantified the HRV predictors of importance using the permutation feature importance structure. This approach is suitable to identify the relationship between the HRV features, task performance, and emotions by identifying a decrease in the machine learning score every time a single HRV feature randomly shuffles [10]. Therefore, we are able to understand how each of these HRV features contributes to the model performance. We applied the permutation feature importance technique for each machine learning model created for both single and two-layer approaches.

3.8 Statistical Analysis

Statistical significance was determined through repeated measures analysis of variance (RM ANOVA) tests on MAE and RMSE. Significance is reported at $\alpha = 0.05$. The RM ANOVAs were separately run on each model performance metric (MAE, RMSE) to test the effects of the two independent variables, emotions (emotions/no emotions) and model type (SVR, MLPR, DTR). Separate RM ANOVAs were run on the HRV and RRV importance to test the effect of the four independent variables: emotions, algorithm type, sex, and HRV, RRV feature importance. Post hoc comparisons were performed where needed using Tukey HSD-Test.

4 RESULTS

4.1 Emotional and Cognitive Workload Validation

4.1.1 Valence and Arousal Validation.

To identify the effectiveness of the music videos to elicit the desired affective states, we ran a paired-samples t-test on the ground truth levels of valence and arousal of the DEAP dataset and the mean reported levels of our experiment. The valence results from the ground truth DEAP dataset (mean=5.33, SD=0.87) and our experiment (mean=5.20, SD=1.10) suggested that the valence levels did not vary significantly (t(22)=0.33, p-value = 0.739). Similarly, arousal results from the ground truth DEAP dataset (mean=4.99, SD=1.67) and our experiment (mean=4.57, SD=0.94) suggested that the arousal levels did not vary significantly (t(22)=0.76, p-value=0.457). These findings suggested that the emotions elicited from the DEAP dataset did not vary significantly from those we generated in the lab. Based on these findings, we used the participants' self-assessments as labels in our dataset.

4.1.2 Cognitive Workload Validation.

A paired-samples t-test was conducted to compare the TLX- Score collected after the low cognitive workload task vs. high cognitive workload task. The results from the low cognitive workload task (mean=37.9, SD=14.37) and high cognitive workload task (mean=52.2, SD=14.5) suggested that the cognitive workload required for completing two tasks varied significantly (t(27)=-6.68, p-value<0.001). Similarly, participants reported significantly (p-value < 0.05) higher mental demand, temporal demand, and effort during the high workload task. To validate the effect of cognitive load on user performance, we compared the task performance for the low cognitive workload task vs. the high cognitive workload task. The results from the low cognitive workload task (mean=33.2, SD=6.5) and high cognitive workload task (mean=42.9, SD=12.8) suggested that the task performance for these two tasks varied significantly (t(27)=-5.59, p-value<0.001). Here, a high score corresponds to lower performance as the task performance measures the deviation from the actual point in pixels. Based on these findings we decided to utilize task performance as the prediction label. To calculate task performance, the Root Mean Square Deviation from the Center point in pixel units was used as the label in our dataset.

4.2 Emotional Assessment Models Performance

Table 1 represents the mean and standard deviation of the MAE and RMSE for all participants. On analyzing the evaluation metrics, the SVR outperformed the DTR and RFR at predicting arousal. The mean MAE and RMSE of the LOSO cross-validation for all the participants were 1.35 and 3.09 respectively. The DTR performed better than the RFR and achieved mean MAE and RMSE of 1.42 and 3.2 respectively. Finally, RFR achieved the highest mean MAE and RMSE of 1.45 and 3.37 respectively. For predicting the valence the DTR algorithm achieved the lowest mean MAE and RMSE of 1.13 and 2.16. The next best algorithm for predicting valence was SVR and achieved mean MAE and RMSE of 1.49 and 3.3. Finally, the RFR achieved mean MAE and RMSE of 1.52 and 3.58 respectively. Based on these results we chose the SVR model for emotional arousal and the DTR for emotional valence assessment. Imperatively, the next step was to connect the first layer (EAM) to the second layer of the machine learning task performance prediction model.

4.3 Task Performance Prediction

Table 2 represents the results of the two-layer and single-layer models. For the two-layer approach, the SVR was the best performing algorithm followed by MLPR and DTR was the worst. In the case of the single-layer approach, the MLPR algorithm outperformed the SVR and DTR. For both models, DTR was the worst-performing algorithm. To identify if adding emotions had a significant impact on the task prediction performance we compared the single-layer models to the two-layer approach.

In comparing the task performance prediction results between single layer model to the two layer approach, no significant differences were observed for the main effect of model type (p-value=0.090, η^2=0.03), and emotions (p-value=0.268, η^2=0.01) on the dependent variable MAE. Further, we did not observe any interaction effect of model type and emotions (p-value=0.634, η^2=0.01) on

Table 1: Emotion Assessment Model Results

Model	Arousal		Valence	
	MAE	RMSE	MAE	RMSE
Decision Tree Regressor	1.42 ± 0.44	3.2 ±1.78	1.13 ± 0.34	2.16 ± 1.35
Random Forest Regressor	1.45 ± 0.41	3.37±1.77	1.52 ± 0.43	3.58 ± 1.99
Support Vector Regressor	1.35 ± 0.43	3.09 ± 1.78	1.49 ± 0.45	3.30 ± 2.02

the dependent variable MAE. However, we observed a significant difference for the main effect of model type (p-value=0.046, η^2=0.04) on the dependent variable RMSE. But, no significant main effect difference was observed in the case of emotions (p-value=0.33,η^2=0.01). Post hoc comparisons using Tukey HSD test indicated that DTR (mean RMSE=98.8, SE=11.1) varied significantly (p-value=0.04) from SVR (mean RMSE=72.05, SE=9.18) and from MLPR (mean RMSE=68.49, SE=7.18, p-value=0.02).However, no significant differences (p-value=0.78) were observed between MLPR and SVR. MLPR had the lowest RMSE suggesting it was the best model followed by SVR and DTR. On investigating interaction effects, we did not observed any interaction effect of model type and emotions (p-value=0.681, η^2=0.01) on the dependent variable RMSE.

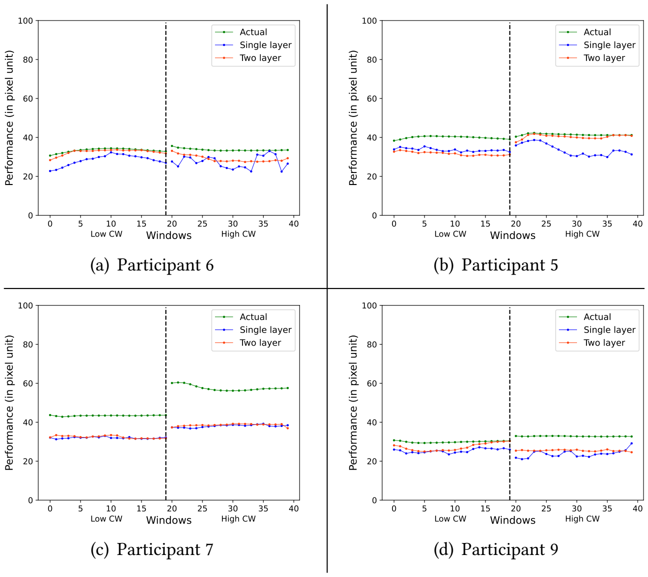
For the two-layer on analyzing the individual performance achieved using SVR algorithms, two participants had MAE values beyond three standard deviations, and five participants had RMSE values beyond three standard deviations. Irrespective of this, SVR had the lowest MAE and RMSE, suggesting that this algorithm could predict well in most cases. For both MLPR and DTR, no participants had MAE values beyond three standard deviations. Finally, for DTR, one participant had an RMSE value beyond three standard deviations. Although there were no outliers, both MLPR and DTR had a higher MAE and RMSE, suggesting that these algorithms did not predict as well as SVR.

4.3.1 *Comparisons and Differences Across single-layer and two-layer Models.*

Although, no significance was observed for independent variable emotions and their interaction effects with the model type we performed a detailed descriptive analysis to understand what impact emotions had on model performance. One of the preliminary observations was that irrespective of the model, the predictions were better by adding emotional information as seen in Table 2. On further investigating the impact of emotions, we observed that when emotion information were added to the model, SVR had a lower MAE (n=28, mean MAE=6.07, SE=0.62) and RMSE (n=28, mean RMSE=60.4, SE=12.3) compared to MLPR MAE (n=28, mean MAE =6.55, SE = 0.61) and RMSE (n=28, mean RMSE=64.76, SE=9.89) and DTR MAE (n=28, mean MAE=7.92, SE=0.65) and RMSE (n=28, mean RMSE=98.5, SE=15.7) suggesting that SVR performed the best when emotion information was added. For the case where information regarding emotions were not added, MLPR had a lower MAE (n=28, mean MAE=7.01, SE=0.62) and RMSE (n=28, mean RMSE=72.2, SE=10.5) compared to SVR MAE (n=28, mean MAE=7.29, SE=0.60) and RMSE (n=28, mean RMSE=83.7, SE=13.5) and Model 3 MAE (n=28, mean MAE=7.96, SE=0.66) and RMSE (n=28, mean RMSE=99.1, SE=16.0) suggesting that SVR performed the best without the emotion information.

4.3.2 *Graph Representation Comparison of the Single and two-layer Models on Some Representative Participants.*

Additionally, to compare the ability of the best performing model (SVR) in predicting the task performance we present a few individual scenarios where the two models (single-, and two-layer) perform differently. The specific scenarios represented the variability among individuals that plays a significant role in the model's efficacy. Specifically, we represent four cases: 1) where the two-layer model performed better at predicting the low cognitive workload state, 2) where the two-layer model did a better job at predicting the high cognitive workload state, 3) where the two models behaved similarly at predicting the task performance, 4) where the two-layer model performed better at predicting the task performance. In each graph, the y-axis represents the task performance in pixel units, and the x-axis is the prediction window for the high and low cognitive workload states. The green dotted line shows the actual task performance of the user. The orange and the blue show the absolute difference of the predicted performance from the actual performance of the two-layer and single-layer models respectively.



(a) Participant 6

(b) Participant 5

(c) Participant 7

(d) Participant 9

Figure 3: Comparing Task Performance Prediction of the Single Layer and Two Layer model

The Figure 3 (a) shows the performance of the participant 6. In this case, it can be observed that the two-layer model is very accurate and performs better than the single layer for all the prediction windows for the low cognitive workload state. In the high

Table 2: Task Performance Prediction Results

Model	Single Layer		Two Layer	
	MAE	RMSE	MAE	RMSE
Support Vector Regressor	7.30 ± 3.22	83.67 ± 71.71	6.07 ± 3.28	60.42 ± 65.12
Desicion Tree Regressor	7.96 ± 3.51	99.12 ± 84.74	7.92 ± 3.47	98.47 ± 83.30
Multilayer Perceptron Regressor	7.02 ± 3.31	72.21 ± 55.76	6.55 ± 3.27	64.75 ± 52.30

cognitive workload state, the two-layer model performed better than the single-layer model for the majority of the prediction windows. Next, we present a case (Figure 3(b)) where the two-layer performed very well at predicting the task performance in the high cognitive workload level wherein all windows the two-layer model was better than the single-layer models and very close to the actual performance. However, in the low cognitive workload state, the single-layer outperformed the two-layer model for all the prediction windows. Participant 7 did not perform well compared to the other participants. Figure 3(c) demonstrates the performance of the two models for this particular participant for the two cognitive workload states.Both models performed almost the same on this participant with the two-layer model outperforming the single-layer model at some predicting windows. Finally, Figure 3 (d) shows a case where the two-layer model performs better than the single-layer models in both low and high cognitive states for almost all the prediction windows. In both cases, the absolute difference between the actual performance and the two-layer prediction performance is smaller in the low than the high cognitive workload state. For only two windows the single-layer models perform better than the two-layer model in the high cognitive workload state.

4.4 Significant Predictors

Irrespective of emotions added or not, we observed that all model types had SD RSP among their top five important features. Although there we no other overlapping features between the three model types, for the two best performing models (i.e., SVR, MLPR), we observed RRV SDBB to be an overlapping feature and was the first and second important feature. Similarly, across MLPR and DTR, we observed two overlapping features: HR mean and RSP mean, of which HR mean was ranked higher in both models. However, more interestingly, we observed that for SVR, Valence and Arousal (features associated with emotions), were among the top five features, which further explains why SVR was the best performing model when emotion details were added as opposed to MLPR which was the best performing model without emotion details. Additionally, for SVR, among the top five features, three are respiration signal features, and two are emotional features with no heart rate signal features. While for MLPR and DTR, we observed only heart rate signal features and respiration signals among the top five features.

5 DISCUSSION

In this paper, we developed three machine learning regression models for predicting the user's task performance during cognitive tasks. We developed an emotional assessment model that can capture the emotional state of the user while performing the tasks and we added this information to another model along with physiological data to predict the users' task performance.

The key takeaways of the study include:

(1) Model performance improves after adding emotional information irrespective of the model type.
(2) Information gained from the emotional assessment output is not linear.
(3) Model type and significant predictors are important to build robust models.

5.1 Model Performance Improves After Adding Emotion Information Irrespective of Model Type

Previous research has reported that cognitive workload influences emotions and physiological responses [19, 50, 52]. However, these relationships were explored independently. This study which developed a novel two-layer approach that captures the interplay of emotions and the physiological state of the user observed that irrespective of learning algorithms two-layer approach leads to more accurate task performance predictions compared to the independent single-layer approach. These findings suggest that when trying to predict task performance the independent variables (i.e., emotions, physiological responses, cognitive workload) should be considered simultaneously rather than independently.

5.2 Information Gain from the Emotional Assessment Output is Not Linear

The information gained by adding the emotions to the two-layer approach was evident where all two-layer models outperformed their respective single-layer models in predicting task performance. However, it should be noted that the information gained by each model is non-linear; wherein the initial single-layer approach, the best performing model was the MLPR, but in the two-layer method, the SVM outperformed MLPR. Additionally, we notice that decision trees did not gain much information from emotion. However, the two algorithms made significant improvements which clearly suggests that each algorithm uses the information in a different manner to make predictions. SVR works by fitting the best line within a specific set threshold value which represents the distance between the hyperplane and boundary line. By adding valence and arousal points the SVR was able to fit a better line in the hyperplane to better predict the task performance. We will add this talking point to our discussion section. This finding suggests that information gain was higher in SVM than MLPR by adding the emotional assessment model, which begs the question of how to choose suitable learning algorithms to develop robust machine learning predictors.

5.3 Model Type and Significant Predictors are Important to Build Robust Models

The statistical analysis showed that among the top five predictors for each model there was only one common feature (SD RSP). Additionally, the emotional-related features (i.e., valence and arousal) are among the top predictors only in the case of SVR. These results indicate that the choice of algorithm type and features is very important to developing models that can capture the underlying relationship between cognitive workload and emotional state to predict task performance. In the case of SVR, the underlying rule is to find the hyperplane that has the maximum number of points meaning that the algorithm relies a lot on the geometrical properties of the data. Our results indicate that adding valence and arousal contributed to creating a better boundary thereby improving task prediction.

5.4 Participant variability

There were a few participants whose physiological data were significantly different compared to the group and affected the predictions made by the machine learning models. Although a higher accuracy in prediction could have been achieved by excluding the outlier cases, we decided to include them to replicate a practical scenario where the physiological measures of participants could vary significantly based on various factors. Additionally, including the variable data helped in building robust machine learning models that can be used for future predictions. Although being an in-lab study, we observed variability among participants and this suggests that a higher variability can be expected in a real-world setting. Future studies should consider including these outliers in model training rather than focusing on performance metrics to develop a robust and reproducible model.

5.5 Study Limitations

One of the limitations of this study is the use of the SAM questionnaire to report valence and arousal which can result in partial understanding and reporting valence and arousal perceptions. In this study, we observed a few cases where the two-layer approach performed better at predicting the task performance in the high state and in some cases the low state for the same participant. This is primarily because, in the emotional assessment dataset, there appears to be an under-representation of music videos that induce extreme valence (negative, positive) and high arousal (negative, positive). Therefore the emotional assessment model fails to capture the emotions that represent these extreme arousal and valence values which could improve the information gained to predict the user's task performance. Future work should consider the accuracy of survey responses as the ground truth for someone's emotional state and a representative selection of stimuli that can induce valence and arousal in the whole range. Another limitation is that the participants recruited in this study were college students predominately seeking advanced degrees in engineering. In future work, efforts should be made to ensure a diverse sample of older and younger in the planned participant pool.

6 CONCLUSION AND FUTURE WORK

In this research, we investigated the interplay of emotions and cognitive workload on performance and we proposed a method for predicting users' task performance during a cognitive task. This method considers the user's emotional state during these tasks to enhance the information gained. We evaluated this method on 28 participants, and the user-independent modeling approach showed that this method provides better prediction results than using only physiological data. The most important aspect of this research was to investigate methods for building robust and adjustable user models that can adapt to the emotional state of each individual during a cognitive task to predict task performance.

In recent years, robots have become key elements in achieving manufacturing competitiveness. Especially in industrial environments, such as assembly lines, a strong level of interaction and cooperation is required where humans and robots form a dynamic system that works together towards achieving a common goal or accomplishing a task. However, to ensure the efficiency and productivity of the overall human-robot cooperation, we need to create a collaborative environment where we, the human operators, feel comfortable working with robots and vice versa. A promising way to achieve this is tuning the interaction with the robot depending on the operator's cognitive and emotional state. The models developed in this research will be adopted for the next stage of our research, where an operator performing a collaborative task with a robot will be intervened (velocity adjustments) to achieve the optimal performance in real-time.

ACKNOWLEDGMENTS

We thank all the members of the Human Interaction Lab for helping with the data collection and for the valuable comments raised from discussions with them. Research reported in this publication was supported by the National Science Foundation under award number 1900704. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Science Foundation.

REFERENCES

[1] Hussein A Abbass, William M Mount, D Tucek, and Jean-Philippe Pinheiro. 2011. Towards a code of best practice for evaluating air traffic control interfaces. In *Australian Transport Research Forum, Adelaide, Australia.*

[2] Mustafa Al'Absi, Kenneth Hugdahl, and William R Lovallo. 2002. Adrenocortical stress responses and altered working memory performance. *Psychophysiology* 39, 1 (Jan. 2002), 95–99.

[3] Amy L Alexander, Christopher D Wickens, and David H Merwin. 2005. Perspective and coplanar cockpit displays of traffic information: Implications for maneuver choice, flight safety, and mental workload. *The International Journal of Aviation Psychology* 15, 1 (Nov. 2005), 1–21.

[4] Hasan Ayaz, Ben Willems, B Bunce, Patricia A Shewokis, Kurtulus Izzetoglu, Sehchang Hah, Atul Deshmukh, and Banu Onaral. 2010. Cognitive workload assessment of air traffic controllers using optical brain imaging sensors. *Advances in understanding human performance: Neuroergonomics, human factors design, and special populations* (May 2010), 21–31.

[5] Ashwin Ramesh Babu, Akilesh Rajavenkatanarayanan, James Robert Brady, and Fillia Makedon. 2018. Multimodal approach for cognitive task performance prediction from body postures, facial expressions and EEG signal. In *Proceedings of the Workshop on Modeling Cognitive Processes from Multimodal Data*. 1–7.

[6] Chris Berka, Daniel J Levendowski, Milenko M Cvetinovic, Miroslav M Petrovic, Gene Davis, Michelle N Lumicao, Vladimir T Zivkovic, Miodrag V Popovic, and Richard Olmstead. 2004. Real-time analysis of EEG indexes of alertness, cognition, and memory acquired with a wireless EEG headset. *International Journal of Human-Computer Interaction* 17, 2 (June 2004), 151–170.

[7] Gary G Berntson, J Thomas Bigger Jr, Dwain L Eckberg, Paul Grossman, Peter G Kaufmann, Marek Malik, Haikady N Nagaraja, Stephen W Porges, J Philip Saul, Peter H Stone, et al. 1997. Heart rate variability: origins, methods, and interpretive caveats. *Psychophysiology* 34, 6 (Nov. 1997), 623–648.

[8] Gunilla Bohlin. 1976. Delayed habituation of the electrodermal orienting response as a function of increased level of arousal. *Psychophysiology* 13, 4 (July 1976), 345–351.

[9] Margaret M Bradley and Peter J Lang. 1994. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry* 25, 1 (March 1994), 49–59.

[10] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (Oct. 2001), 5–32.

[11] Per Brodal. 2004. *The central nervous system: structure and function.* Oxford University Press.

[12] John G Casali and Walter W Wierwille. 1984. On the measurement of pilot perceptual workload: a comparison of assessment techniques addressing sensitivity and intrusion issues. *Ergonomics* 27, 10 (May 1984), 1033–1050.

[13] DF Cechetto, CB Saper, AD Loewy, and KM Spyer. 1990. Central regulation of autonomic functions. *Oxford University Press, New York* (1990), 208–223.

[14] Rebecca L Charles and Jim Nixon. 2019. Measuring mental workload using physiological measures: A systematic review. *Applied ergonomics* 74 (Jan. 2019), 221–232.

[15] Zi Cheng, Lin Shu, Jinyan Xie, and CL Philip Chen. 2017. A novel ECG-based real-time detection method of negative emotions in wearable applications. In *Conference on Security, Pattern Analysis, and Cybernetics*.

[16] Burcu Cinaz, Bert Arnrich, Roberto La Marca, and Gerhard Tröster. 2013. Monitoring of mental workload levels during an everyday life office-work scenario. *Personal and ubiquitous computing* 17, 2 (Oct. 2013), 229–239.

[17] Michel De Rivecourt, MN Kuperus, WJ Post, and LJM Mulder. 2008. Cardiovascular and eye activity measures as indices for momentary changes in mental effort during simulated flight. *Ergonomics* 51, 9 (Sept. 2008), 1295–1319.

[18] AK Dey and DD Mann. 2010. A complete task analysis to measure the workload associated with operating an agricultural sprayer equipped with a navigation device. *Applied ergonomics* 41, 1 (Jan. 2010), 146–149.

[19] Mark S Edwards, Philippa Moore, James C Champion, and Elizabeth J Edwards. 2015. Effects of trait anxiety and situational stress on attentional shifting are buffered by working memory capacity. *Anxiety, Stress, & Coping* 28, 1 (April 2015), 1–16.

[20] Stephen H Fairclough and Louise Venables. 2006. Prediction of subjective states from psychophysiology: A multivariate approach. *Biological psychology* 71, 1 (Jan. 2006), 100–110.

[21] Jianzhong Fan, Haihui Li, Yu Zhan, and Yajun Yu. 2019. An electrocardiogram acquisition and analysis system for detection of human stress. In *International Congress on Image and Signal Processing, BioMedical Engineering and Informatics*.

[22] L Finsen, K Søgaard, C Jensen, V Borg, and H Christensen. 2001. Muscle activity and cardiovascular response during computer-mouse work with and without memory demands. *Ergonomics* 44, 14 (Nov. 2001), 1312–1329.

[23] Hannah J Foy and Peter Chapman. 2018. Mental workload is reflected in driver behaviour, physiology, eye movements and prefrontal cortex activation. *Applied ergonomics* 73 (Nov. 2018), 90–99.

[24] Edith Galy. 2018. Consideration of several mental workload categories: perspectives for elaboration of new ergonomic recommendations concerning shiftwork. *Theoretical issues in ergonomics science* 19, 4 (Sept. 2018), 483–497.

[25] Qin Gao, Yang Wang, Fei Song, Zhizhong Li, and Xiaolu Dong. 2013. Mental workload measurement for emergency operating procedures in digital nuclear power plants. *Ergonomics* 56, 7 (May 2013), 1070–1085.

[26] Vishnunarayan Girishan Prabhu, Kevin Taaffe, Ronald Pirrallo, and Dotan Shvorin. 2020. Stress and burnout among attending and resident physicians in the ED: a comparative study. *IISE Transactions on Healthcare Systems Engineering* 11, 1 (Aug. 2020), 1–10.

[27] Alberto Greco, Antonio Lanata, Gaetano Valenza, Giuseppina Rota, Nicola Vanello, and Enzo Pasquale Scilingo. 2012. On the deconvolution analysis of electrodermal activity in bipolar patients. In *Engineering in Medicine and Biology Society*. 6691–6694.

[28] Mark K Greenwald, Margaret M Bradley, Alfons O Hamm, and PJ Lang. 1993. Looking at pictures: evaluative, facial, visceral and behavioral responses. *Psychophysiology* 30, 3 (May 1993), 261–273.

[29] Han-Wen Guo, Yu-Shun Huang, Chien-Hung Lin, Jen-Chien Chien, Koichi Haraikawa, and Jiann-Shing Shieh. 2016. Heart rate variability signal features for emotion recognition by using principal component analysis and support vectors machine. In *Bioinformatics and Bioengineering (BIBE)*. 274–277.

[30] Peter A Hancock and Paula A Desmond. 2001. *Stress, workload, and fatigue.* Lawrence Erlbaum Associates Publishers.

[31] Peter A Hancock and Gerald Matthews. 2019. Workload and performance: Associations, insensitivities, and dissociations. *Human factors* 61, 3 (Dec. 2019), 374–392.

[32] Nis Hjortskov, Dag Rissén, Anne Katrine Blangsted, Nils Fallentin, Ulf Lundberg, and Karen Søgaard. 2004. The effect of mental stress on heart rate variability and blood pressure during computer work. *European journal of applied physiology* 92, 1 (June 2004), 84–89.

[33] G Robert J Hockey. 1997. Compensatory control in the regulation of human performance under stress and high workload: A cognitive-energetical framework. *Biological psychology* 45, 1-3 (March 1997), 73–93.

[34] Eva Hudlicka and Michael D Mcneese. 2002. Assessment of user affective and belief states for interface adaptation: Application to an Air Force pilot task. *User Modeling and User-Adapted Interaction* 12, 1 (Feb. 2002), 1–47.

[35] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. 2011. Deap: A database for emotion analysis; using physiological signals. *Transactions on affective computing* 3, 1 (June 2011), 18–31.

[36] Peter J Lang. 1995. The emotion probe: Studies of motivation and attention. *American psychologist* 50, 5 (May 1995), 372.

[37] Peter J Lang, Margaret M Bradley, Bruce N Cuthbert, et al. 1997. International affective picture system (IAPS): Technical manual and affective ratings. *NIMH Center for the Study of Emotion and Attention* 1, 39-58 (1997), 3.

[38] Ying Lean and Fu Shan. 2012. Brief review on physiological and biochemical evaluations of human mental workload. *Human Factors and Ergonomics in Manufacturing & Service Industries* 22, 3 (Nov. 2012), 177–187.

[39] Gerhard Marquart, Christopher Cabrall, and Joost de Winter. 2015. Review of eye-related measures of drivers' mental workload. *Procedia Manufacturing* 3 (2015), 2854–2861.

[40] Lukasz M Mazur, Prithima R Mosaly, Carlton Moore, and Lawrence Marks. 2019. Association of the usability of electronic health records with cognitive workload and performance levels among physicians. *JAMA network open* 2, 4 (April 2019), e191709–e191709.

[41] Albert Mehrabian. 1997. Comparison of the PAD and PANAS as models for describing emotions and for differentiating anxiety from depression. *Journal of psychopathology and behavioral assessment* 19, 4 (Dec. 1997), 331–357.

[42] Jim Nixon and Rebecca Charles. 2017. Understanding the human performance envelope using electrophysiological measures from wearable technology. *Cognition, Technology & Work* 19, 4 (Sept. 2017), 655–666.

[43] Michalis Papakostas, Konstantinos Tsiakas, Theodoros Giannakopoulos, and Fillia Makedon. 2017. Towards predicting task performance from EEG signals. In *Big Data*.

[44] Rosalind W Picard. 2000. *Affective computing.* MIT press.

[45] Hugo F Posada-Quintero and Jeffrey B Bolkhovsky. 2019. Machine learning models for the identification of cognitive tasks using autonomic reactions from heart rate variability and electrodermal activity. *Behavioral Sciences* 9, 4 (April 2019), 45.

[46] Rodney A Rhoades and David R Bell. 2012. *Medical phisiology: Principles for clinical medicine.* Lippincott Williams & Wilkins.

[47] Kenneth S Saladin. 2005. *Human anatomy.* Rex Bookstore.

[48] Fred Shaffer and Jay P Ginsberg. 2017. An overview of heart rate variability metrics and norms. *Frontiers in public health* (2017), 258.

[49] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. 2011. A multimodal database for affect recognition and implicit tagging. *IEEE transactions on affective computing* 3, 1 (Aug. 2011), 42–55.

[50] Barbara A Sorg and Paul Whitney. 1992. The effect of trait anxiety and situational stress on working memory capacity. *Journal of research in personality* 26, 3 (April 1992), 235–241.

[51] Gaetano Valenza, Luca Citi, Antonio Lanatá, Enzo Pasquale Scilingo, and Riccardo Barbieri. 2014. Revealing real-time emotional responses: a personalized assessment based on heartbeat dynamics. *Scientific reports* 4, 1 (May 2014), 1–13.

[52] Bram B Van Acker, Davy D Parmentier, Peter Vlerick, and Jelle Saldien. 2018. Understanding mental workload: from a clarifying concept analysis toward an implementable framework. *Cognition, technology & work* 20, 3 (Aug. 2018), 351–365.

[53] Christopher D Wickens. 2002. Multiple resources and performance prediction. *Theoretical issues in ergonomics science* 3, 2 (Nov. 2002), 159–177.

[54] Wickens, Christopher D and McCarley, Jason S. 2019. *Applied attention theory.* CRC press.

[55] Glenn F Wilson and Christopher A Russell. 2003. Real-time assessment of mental workload using psychophysiological measures and artificial neural networks. *Human factors* 45, 4 (Dec. 2003), 635–644.

[56] Mark S Young and Neville A Stanton. 2002. Attention and automation: new perspectives on mental underload and performance. *Theoretical issues in ergonomics science* 3, 2 (2002), 178–194.