Classifying Cognitive Workload Using Machine Learning Techniques and Non-Intrusive Wearable Devices

Yunmei Liu
Department of Industrial Engineering
University of Louisville
Louisville, KY USA
Email address:
yunmei.liu@louisville.edu

Nicolas S. Grimaldi
Department of Industrial & Systems Engineering
University of Florida
Gainesville, FL USA
Email address:
n.grimaldi@ufl.edu

Niosh Basnet
Wm Michael Barnes '64
Department of Industrial & Systems Engineering
Texas A&M University
College Station, TX USA
Email address:
niosh@tamu.edu

David Wozniak
Wm Michael Barnes '64
Department of Industrial & Systems Engineering
Texas A&M University
College Station, TX USA
Email address:
ace1208@tamu.edu

Eric Chen
Department of Computer & Information Science & Engineering
University of Florida
Gainesville, FL USA
Email address:
eyc14@duke.edu

Maryam Zahabi
Wm Michael Barnes '64
Department of Industrial & Systems Engineering
Texas A&M University
College Station, TX USA
Email address:
mzahabi@tamu.edu

David Kaber
Department of Industrial & Systems Engineering
University of Florida
Gainesville, FL USA
Email address:
dkaber@ufl.edu

Jaime Ruiz
Department of Computer & Information Science & Engineering
University of Florida
Gainesville, FL USA
Email address:
jaime.ruiz@ufl.edu

***Abstract*—**Effective management of cognitive workload is essential to ensure user performance and minimize mental fatigue. This study aimed to evaluate the effectiveness of three machine learning (ML) algorithms, including a Support Vector Machine (SVM), a Random Forest (RF) and Random Fourier Features (RFF) models, for predicting instances of high cognitive workload based on physiological data collected from non-intrusive wearable devices. The study included scenarios to induce high workload. After processing data from 30 human participants, and extracting cognitive workload-related features, we evaluated the ML algorithms in terms of accuracy, Area Under the Curve (AUC), and Average Precision (AP) responses. The RFF model emerged as a top performer, achieving 97% accuracy, 99.5% AUC, and 99% AP, demonstrating the capability to detect high workload scenarios and robustness to unbalanced data in workload predictions. The most significant physiological features in the RFF model included skin conductance level (SCL) change, SCL mean, and percent change in pupil size. Future studies should enhance the present binary classification model to support multi-class or continuous workload predictions and implement real-time analysis using trained algorithms .

***Keywords*—** Cognitive workload, machine learning, physiological signals, workload prediction

I. Introduction

Elevated cognitive workload has been correlated with diminished operator task accuracy [1] and increased mental fatigue [2]. In industries such as commercial transportation, operator attention and cognitive workload indicators, including gaze behavior, have been found to be useful for predicting driving task performance [3]. However, in other industries, such as healthcare, there is a gap or absence of adequate data-based systems to assess and facilitate cognitive task allocation. This situation can lead to increased workload, higher task error rates, and worker frustration [4]. Monitoring and addressing cognitive workload is essential, especially when operators undertake tasks of high difficulty that pose considerable demands on cognitive resources/processes and can adversely affect concentration and lead to errors[5].

Recently, there has been a surge in developing methods for classifying and predicting states of cognitive workload. Machine learning (ML) algorithms, including multi-dimensional scaling and high-order tensor decompositions, have been used to identify complex structures and patterns in high-dimensional datasets [6]. Most current approaches use neurophysiological, behavioral, and/or physiological measures to classify cognitive workload states. Neurophysiological measures offer the most direct reflection of cognitive workload states but can also be intrusive to task performance and impractical in field operations [7]. In contrast, behavioral measures are effective but do not provide consistent and frequent data rapidly enough for continuous real-time workload monitoring. Physiological measures combine the benefits of both offering continuous data streams and non-invasive measurement devices [8].

The emergence of multimodal sensor analysis and integration of data from various on-body sensors, has proven to be the preferred methodology, demonstrating distinct advantages over single-sensor analyses [7]. Among physiological signals, those considered applicable for real-time data acquisition in field activities are galvanic skin response (GSR) [9, 10], blood volume pulse (BVP) [11,12], and eye-tracking [13]. Each of these signals have features correlating with stress and workload responses. For GSR, the skin conductance response (SCR; half-time, recovery time, and amplitude) as well as skin conductance level (SCL; mean and standard deviation) have been shown to be meaningful indicators of stress states [9,10]. Regarding BVP, mean amplitude, the ratio of low frequency / high frequency (LF/HF) components of heart rate variability (HRV) [11,12], and inter-

979-8-3503-1579-0/24/$31.00 ©2024 IEEE

beat interval (IBI; mean and standard deviation) [11] are some of the most common features used as workload indicators. Finally, eye-tracking measures, including blink rate and percent change in pupil size (PCPS), have been found to be notable measures linked with cognitive workload [13].

Utilizing these data streams, several ML classification models have been developed to distinguish varying levels of cognitive workload, stress, and emotions, including Support Vector Machines (SVM) [14-16], Random Forest (RF) models [16,17], and Random Fourier Features (RFF) [16]. In addition, meticulous compilation of a robust set of input features is critical to attaining superior classification accuracy [18].

GSR measurements can be collected using wearable devices such as the Empatica E4 [19], Moodmetric Ring [20], and Empatica Embrace 2 [21]. For BVP measurements, devices like the Polar OH1 [22], Polar H10 [22], and Samsung Gear S2 [22] have been used in prior studies along with the Empatica E4 [19]. The Empatica device has dual capacity to capture both GSR and BVP signals. For eye tracking, the Pupil Labs' Pupil Core [23] and Tobii Pro Glasses [24] are at the forefront of workload assessment due to their small sizes and light weights.

The primary objective of this study was to assess the effectiveness of physiological measures collected with non-obtrusive devices, including the Empatica E4 and Pupil Labs eye tracking glasses, for classifying cognitive workload states. Additionally, we sought to extract high-quality features from the multimodal physiological data and to evaluate the impact on classification model performance. Lastly, by testing a set of suitable ML algorithms, we aimed to identify the highest performing model and to set a benchmark for cognitive workload classification.

II. Method

A. Participant and Apparatus

Thirty participants (17 males, 13 females) were recruited for this study (Age: M= 26.57 yrs.; SD = 5.72 yrs.). All participants had 20/20 vision and no hearing impairment. The experiment protocol was approved by the Institutional Review Board at Texas A&M University.

Fig. 1. (left) Empatica E4 wristband; (right) Eye-tracking device.

Two types of wearable devices were used in the study, including the Pupil Labs eye-tracking glasses and an Empatica E4 wristband (Figure 1). The Pupil Labs eye-tracking system is a lightweight head-mounted and low power wired device with up to 200 Hz sampling frequency that measures blink rate and pupil diameter. The Empatica E4 wristband is also a lightweight and low power device worn (untethered) on the non-dominant wrist of a user that has a 20-36-hour battery life and measures GSR and BVP at varying frequencies.

B. Experiment Design

For this experiment, we simulated cooking tasks in which multiple subtasks must be addressed simultaneously under time constraints or when unexpected events occur. These circumstances can create moments of high workload, closely resembling other safety critical \ tasks posing high cognitive demands. Through accurate detection of instances of high cognitive load in cooking tasks, it may be possible to design timely task assistance to help operators navigate through stressful periods. With this in mind, we designed an experiment to collect operators' physiological signals in the context of cooking. The experiment involved three cooking recipes, including: pinwheel appetizers, pour-over coffee, and (coffee) mug cakes.

From results of a pilot test, we determined that a majority of task time consisted of low workload operations. This was attributed to the fact that cooking tasks are generally not cognitively demanding. Given this challenge, we redesigned the experiment to capture more instances of high workload by artificially introducing task time pressure and error conditions.

The full experiment followed a within-subject design. Participants performed under both nominal and off-nominal task conditions for each recipe. In the nominal condition, they completed recipes without any time constraint or externally induced errors. In contrast, in the off-nominal condition, participants were subjected to a tight task time constraint and circumstances were presented to induce errors, simulating high workload levels in recipe preparation. The time pressure and error manipulations were custom designed for each recipe. In addition, each recipe began with the off-nominal condition, ensuring that participants' initial exposure to the recipe was unprepared, which in turn elevated their workload. At the end of each step, participants rated their workload level subjectively on a scale from 1 (very low) to 5 (very high), as a means for labeling the task trials in terms of perceived workload.

C. Study Procedure

When participants arrived at the lab, they signed an informed consent form and completed a demographic survey. They were subsequently briefed on the experiment objectives and procedures. They were then outfitted with the eye-tracking device and an Empatica wristband was fastened to the wrist of their non-dominant hand. Participants adjusted the wristband for comfort.

For eye tracking system calibration, markers were placed on each corner of the physical (cooking) workstation. Participants were instructed to remain stationary in front of the table, and to maintain a fixed head position. They were then asked to concentrate on the farthest left-side marker and then repeat this action in a counterclockwise direction for all markers. Subsequently, baseline gaze measurements were taken as participants looked at a white cross image against a black background on a wall in front of them for a duration of 5 min.

Upon completion of the baseline data collection, each participant performed 12 unique test trials. For each recipe, participants completed four trials with the first two being

training sessions (under nominal task conditions) while the subsequent two were for experimental data collection on nominal and off-nominal task performance. At the end of the experiment, baseline measurements were recorded once again. The entire experiment took approximately 3 to 3.5 hours varying among participants. For their commitment and effort, participants were compensated at a rate of $15 per hour.

D. Physiological Signals

Research has consistently observed that several physiological responses, such as BVP, GSR, PCPS, and blink rate (BR), have notable correlations with mental states [9-13]. Therefore, we opted to measure GSR and BVP signals, as captured by the Empatica E4, complemented by pupillary and blink data sourced from the PupilLabs eye-tracker.

E. Feature Extraction

Regarding physiological signal processing, feature extraction is fundamental for decoding inherent patterns in data. Due to the inherent noise present in physiological responses, primarily arising from external interference, there is a need for preprocessing of signals prior to extraction of salient features. Our procedure involved three-steps that began with filtering the raw data, data normalization, and then extraction of features from both the time and frequency domains. The NeuroKit2 library, in python, was used to filter and extract features from the GSR and BVP signals [25].

Filtering. Regarding the eye-tracking data, we eliminated any detected blink with a confidence level under 50%, according to the PupilLabs system. Pupil diameter measures with a confidence level under 60% were also disregarded. According to PupilLabs, these confidence levels allow for identification of potentially inaccurate eye response data, while also not compromising the steady stream of accurate (high confidence) data points. BVP was first filtered using a second-order 3Hz lowpass Butterworth filter [26], followed by a third-order bandpass (0.5-8Hz) Butterworth filter [25]. Finally, the GSR signal was filtered using a fourth-order 4Hz lowpass Butterworth filter [25].

Normalization. Pupil size data collected during the baseline trials was used to normalize all test trial observations for feature extraction. Specifically, pupil diameter values were standardized using the mean and standard deviation observed during the baseline sessions.

Feature Extraction. Once all data streams were filtered and normalized, we extracted meaningful features from the datasets. For the GSR data, we distinguished between the SCR and SCL, based on specific frequency components. Features, such as amplitude, rise time and half recovery time were derived from the SCR, while for SCL, we focused on computing metrics, such as the mean value across the signal and its change-over time. Figure 2 presents a visualization of a sample of the GSR feature extraction for a pour over coffee trial without error. The sample is constrained by the Empatica sensor sampling rate.

For the BVP data stream, we focused on identifying the inter-beat interval (IBI) by detecting peaks in the signal BVP. This also led to determination of the BVP amplitude. Furthermore, we derived the ratio between LF/HF components of HRV by ascertaining the power distribution across these frequencies from the BVP data. Similar to Figure 2, Figure 3 presents a visualization of a sample of BVP feature extraction, based on the data from one participant in a nominal task condition trial.

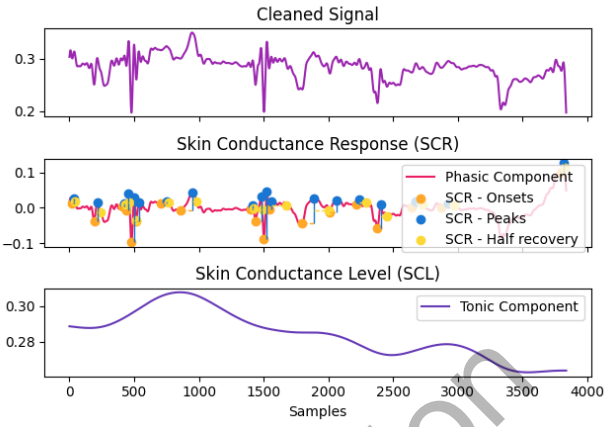

Fig. 2. Processed GSR Signal.

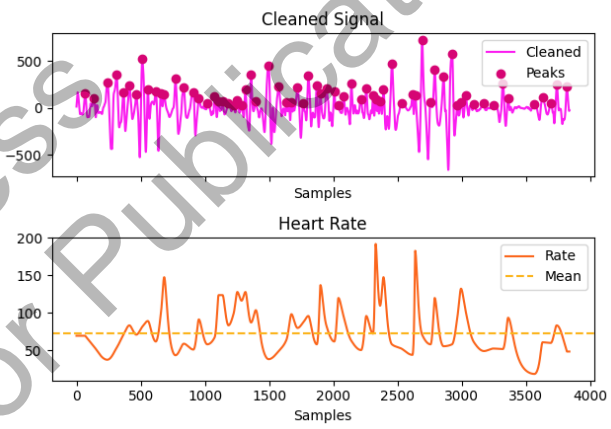

Fig. 3. Processed BVP Signal.

From the pupillary data, we extracted the PCPS feature to provide a basis for inferring cognitive processing speed [27] under the given task conditions. This metric is crucial as it provides information regarding cognitive effort and attention. Lastly, for the eye blink signal, we focused on the BR feature, which serves as an indirect indicator of cognitive load and alertness [28,29].

In the extraction phase, the overarching aim was to gather insights not just as presented but also by leveraging inferential techniques. For instance, while the SCL features were directly computed, some required in-depth analysis like the amplitude, which was discerned from the waveform itself. By referring to the existing literature [9-13], a total of 11 features were selected as having potential utility for our physiological signal analytical framework.

F. Model Training and Evaluation

Model Selection. Considering the prior literature on ML model performance for cognitive workload or stress prediction and computational efficiency [14-18], the SVM, RF and RFF modeling approaches were selected as most suitable for our

study. The RFF algorithm, in particular, stood out due to its capability to approximate Gaussian kernel feature maps and its scalability, speed, and capability to capture non-linear relationships [30].

Data Labeling. In the modeling process, physiological signals were labeled based on participant self-reported cognitive workload. This assessment occurred at every step of a cooking recipe as participants performed the task. The pinwheel and mug cake recipes each included 12 steps. The pour-over coffee recipe consisted of 8 steps. Given that each step was completed within a half-minute, it was presumed that workload remained consistent throughout each individual step. Ratings of 1 and 2 were classified as 'low workload', while ratings of 3 to 5 were classified as 'high workload'. We opted for binary workload state classification in this study because our primary objective was to identify instances of high workload in the task that could compromise performance, as opposed to tracking the trend of workload fluctuations over time. Thus, any physiological response data points gathered during a step that received a rating of 3 to 5 were marked as instances of high workload.

Training. Having curated and preprocessed the experiment data, we implemented a stratified splitting method, allocating 70% of the dataset to training, and 15% each to validation and testing, thus ensuring proportional class distribution across the subsets. This was followed by data normalization and scaling.

Model Evaluation and Optimization. For model evaluation, we employed various performance metrics, including prediction accuracy. The Receiver Operating Characteristic (ROC) curve illustrates a classifiers' diagnostic capability across different thresholds, with the Area Under the Curve (AUC) reflecting a model's discriminative power, where an AUC of 1 indicates perfect discrimination. The Precision-Recall curve is useful for evaluating model performance for unbalanced datasets. The measure pinpoints the balance between precision and model sensitivity. The Average Precision (AP) offers a consolidated model performance score. Together, the accuracy, ROC-AUC and Precision-Recall metrics can provide a comprehensive understanding of model performance in terms of both the quality and precision of predictions. To fine-tune our models for optimal performance and comparisons, we leveraged the Hyperopt library for Bayesian optimization. The objective here was to maximize multiple performance metrics, thus, our tuning metric (or loss function) was computed by minimizing [1 - (Accuracy + F1 + Precision + Recall + AUC)/5]. The F1 measure combines model precision and recall scores and is calculated as the harmonic mean of precision and recall.

Feature Importance. In this study, we examined physiological signal features relevant to cognitive workload. However, device limitations necessitated a closer investigation into the importance of each feature to support accurate model predictions. We employed the SHAP (SHapley Additive exPlanations) method to produce values of the impact of each feature in model predictions. The approach essentially highlights the difference in ML mnodel outcomes when a particular feature is excluded as a model input [33].

III. Results

The dataset from the experiment included a minor imbalance between 'Low' and 'High' cognitive workload levels, with 'Low' comprising 64.95% (30,082 instances) and 'High' comprising 35.05% (16,022 instances) across all 180 test trials. The RFF achieved the highest accuracy of 97.4%, followed closely by RF at 93.1%, and then SVM at 85.7%.

In Figure 4, the ROC curves visualize the trade-offs between the True Positive Rate (TPR) and False Positive Rate (FPR) for the three models. Among the classifiers examined, the RFF classifier distinctly stood-out with an AUC of 0.995. It demonstrated good precision, correctly identifying 96.8% of true high workload instances (TPR = 0.968), while misclassifying low workload instances as high workload in only 2.1% of observations (FPR = 0.021). Following closely was the RF classifier, which registered an AUC of 0.991 with a TPR of 0.964 and an FPR of 0.056. In comparison, the SVM classifier, with an AUC of 0.920, trailed behind the RFF and RF classifiers. Although it accurately predicted high workload instances in 82.5% of observations (TPR = 0.825), it misclassified low workload instances as high workload for 15.3% of observations (FPR = 0.153).

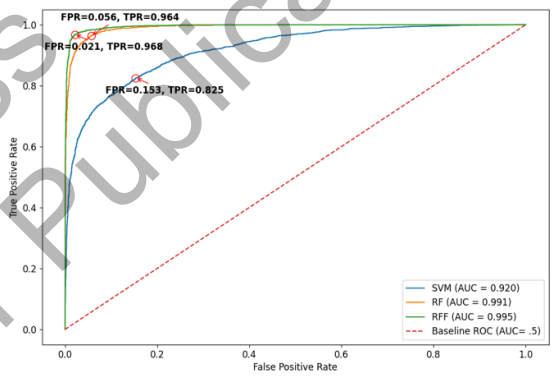

Fig. 4. ROC curves for three models.

Figure 5 presents the Precision-Recall curves, highlighting the interplay between model precision and sensitivity for the classifiers under consideration. These curves are particularly insightful in scenarios with unbalanced datasets, as they focus more on the performance of the model with respect to the positive (minority) class. The RFF classifier performed the best with an AP of 0.991. A closer examination of the PR curves revealed that the RFF classifier delivered a high precision of 0.968 (i.e., when the classifier predicted high workload, it was correct 96.8% of the time), even at an elevated recall level of 0.939 (i.e., the classifier identified 93.9% of all the actual instances of high workload). Following closely, the RF classifier exhibited an AP of 0.985, with a precision of 0.937 at a similar recall of 0.939. However, the SVM classifier trailed with an AP of 0.884. It demonstrated a precision of 0.815 (i.e., when the classifier predicted high workload, it was correct 81.5% of the time) at a recall of 0.763 (i.e., the classifier identified 76.3% of all the actual instances of high workload), indicating a slight drop in its capability to correctly identify positive instances as the recall increased. For reference, the baseline, representing the proportion of positive instances in the dataset, had an AP of 0.350.

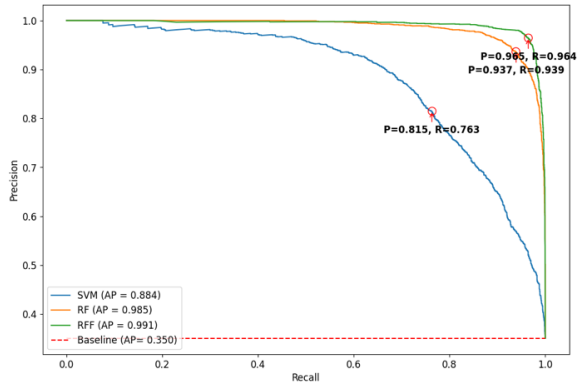

Fig. 5. Precision-Recall curves for three models.

Figure 6 highlights the importance of each feature in contributing to the RFF model's predictions through SHAP values. All features contributed to the model's performance. The most significant features were SCL change, SCL mean, and PCPS value. Opposite to these features, the SCR rise time, SCR recovery time, and the LF/HF ratio had limited importance for RFF model performance.

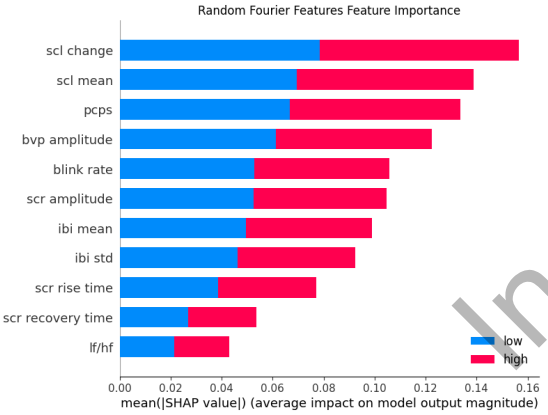

Fig. 6. RFF Features Importance.

IV. Discussion and Conclusion

The primary goal of this study was to evaluate the effectiveness of various ML models in predicting cognitive workload levels, using data gathered from non-intrusive/on-body physiological response sensors, and to achieve high model performance. Through multiple evaluation methods, including accuracy, AUC and AP, we found that the RFF model exhibited the highest performance, resulting in a 0.974 accuracy, 0.995 AUC, and 0.991 AP. This indicates that the RFF model is not only sensitive for detecting high workload scenarios but is also robust for handling unbalanced data when predicting high workload instances. These results surpass those reported in similar prediction research in the existing literature, encompassing both real-time and offline cognitive workload predictions [10,15-18,34].

Despite tailoring the experiment design to balance subject exposure to low and high workload conditions, we also further explored hybrid synthetic methods for balancing observations, including the Synthetic Minority Oversampling Technique (SMOTE) and Edited Nearest Neighbors (ENN) method with 'n_neighbors=3'. This combination leverages the benefits of oversampling and undersampling to improve predictive accuracy in datasets with few positive instances [31]; i.e., high workload, in our case. The superior performance of SMOTE-ENN over either SMOTE or ENN alone, as highlighted in other studies [e.g., 32], further justified its inclusion of the approach in our data preprocessing phase. However, upon evaluating the models, we observed that all synthetically balanced datasets did not notably enhance model performance. As a result, we chose not to report corresponding findings.

Our study employed several strategies to optimize model performance. Initially, we refined the experiment design to collect a relatively balanced and high-quality dataset. Our attempts to further balance the dataset using SMOTE-ENN revealed the synthetic methods to address disparities between observations on 'Low' and 'High' workload but did not significantly enhance model performance. This finding accentuates the inherent quality of our original dataset, suggesting that it already possessed meaningful patterns and structures essential for ML model training. In addition, our data preprocessing and feature extraction methodologies allowed us to extract high-quality, pertinent features from the signals captured by wearable devices, thereby contributing to the model's efficacy. Our feature importance analysis further illuminated the significance of specific physiological signal features for predicting cognitive workload. The literature review guided our selection of widely used algorithms for cognitive workload prediction and provided a useful platform for improving on prior study results.

The capability to accurately identify instances of high cognitive workload in task performance by using non-intrusive wearable devices can provide a basis for control measures to ensure that task performance goes uninterrupted and without errors. This capability may be especially important in extreme situations characterized by time constraints or unforeseen safety issues, where support and intervention (informed by precise workload assessments) can significantly reduce user and system risks. It is important to highlight that while our study focused on cooking tasks, the documented approach is versatile and can be adapted to various other task scenarios.

Although our ML model achieved high performance, there are some limitations to be addressed in future studies. The current binary classification approach identified high and low workload levels. Future studies could classify multiple levels of workload or make continuous predictions to represent trends in workload. Our results were based on a singular dataset. To ascertain the robustness and generalizability of RFF and similar ML models, further research involving a variety of benchmark datasets is recommended. Additionally, the primary focus of this study was to compare the relative performance of different models rather than pursuing the optimization of their absolute performance. Furthermore, our current model evaluation was conducted offline; however, we plan to use the trained algorithm for real-time workload prediction in a future study. Lastly, the models, initially trained on cooking-related tasks, present an opportunity for expansion to encompass a broader range of activities, particularly those involving complex task guidance under high workload conditions.

ACKNOWLEDGMENT

This work was partially funded by the U.S. Defense Advanced Research Projects Agency under contract #HR00112220004. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect these agencies' views.

REFERENCES

[1]. Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. Human factors, 39(2), 230-253.

[2]. Irfan, M. (2022). Measurement of Mental Workload and Fatigue of Production Operator. International Journal of Service Science, Management, Engineering, and Technology, 1(3), 11-13.

[3]. Tsai, Y. F., Viirre, E., Strychacz, C., Chase, B., & Jung, T. P. (2007). Task performance and eye activity: predicting behavior relating to cognitive workload. Aviation, space, and environmental medicine, 78(5), B176-B185.

[4]. National Research Council. (2009). Computational technology for effective health care: immediate steps and strategic directions.

[5]. Becerra-Sánchez, P., Reyes-Munoz, A., & Guerrero-Ibañez, A. (2020). Feature selection model based on EEG signals for assessing the cognitive workload in drivers. Sensors, 20(20), 5881.

[6]. Moustafa, K., Luz, S., & Longo, L. (2017). Assessment of mental workload: a comparison of machine learning methods and subjective assessment techniques. In Human Mental Workload: Models and Applications: First International Symposium, H-WORKLOAD 2017, Dublin, Ireland, June 28-30, 2017, Revised Selected Papers 1 (pp. 30-50). Springer International Publishing.

[7]. Debie, E., Rojas, R. F., Fidock, J., Barlow, M., Kasmarik, K., Anavatti, S., ... & Abbass, H. A. (2019). Multimodal fusion for objective assessment of cognitive workload: a review. IEEE transactions on cybernetics, 51(3), 1542-1555.

[8]. Ranchet, M., Morgan, J. C., Akinwuntan, A. E., & Devos, H. (2017). Cognitive workload across the spectrum of cognitive impairments: A systematic review of physiological measures. Neuroscience & Biobehavioral Reviews, 80, 516-537.

[9]. Makowski, D., Pham, T., Lau, Z. J., Brammer, J. C., Lespinasse, F., Pham, H., Schölzel, C., & Chen, S. A. (2021). NeuroKit2: A Python toolbox for neurophysiological signal processing. Behavior Research Methods, 53(4), 1689–1696.

[10]. Swangnetr, M., & Kaber, D. B. (2013). Emotional state classification in patient–robot interaction using wavelet analysis and Statistics-based feature selection. IEEE Transactions on Human-Machine Systems, 43(1), 63–75.

[11]. Alberdi, A., Aztiria, A., & Basarab, A. (2016). Towards an automatic early stress recognition system for office environments based on multimodal measurements: A review. Journal of Biomedical Informatics, 59, 49–75.

[12]. Shaffer, F., & Ginsberg, J. P. (2017). An Overview of Heart Rate Variability Metrics and Norms. Frontiers in public health, 5, 258.

[13]. Koskinen, J., Bednarik, R., Vrzakova, H., & Elomaa, A. P. (2020). Combined Gaze Metrics as Stress-Sensitive Indicators of Microsurgical Proficiency. Surgical innovation, 27(6), 614–622.

[14]. Chen, L. L., Zhao, Y., Ye, P. F., Zhang, J., & Zou, J. Z. (2017). Detecting driving stress in physiological signals based on multimodal feature analysis and kernel classifiers. Expert Systems with Applications, 85, 279-291.

[15]. Cosoli, G., Iadarola, G., Poli, A., & Spinsante, S. (2021, June). Learning classifiers for analysis of Blood Volume Pulse signals in IoT-enabled systems. In 2021 IEEE International Workshop on Metrology for Industry 4.0 & IoT (MetroInd4. 0&IoT) (pp. 307-312). IEEE.

[16]. Zahabi, M., Wang, Y., & Shahrampour, S. (2021). Classification of officers' driving situations based on eye-tracking and driver performance measures. IEEE Transactions on Human-Machine Systems, 51(4), 394-402.

[17]. Luong, T., Martin, N., Raison, A., Argelaguet, F., Diverrez, J. M., & Lécuyer, A. (2020, November). Towards real-time recognition of users mental workload using integrated physiological sensors into a VR HMD. In 2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR) (pp. 425-437). IEEE.

[18]. Kukolja, D., Popović, S., Horvat, M., Kovač, B., & Ćosić, K. (2014). Comparative analysis of emotion estimation methods based on physiological measurements for real-time applications. International journal of human-computer studies, 72(10-11), 717-727.

[19]. E4 wristband: Real-time Physiological Signals: Wearable PPG, EDA, Temperature, Motion Sensors. Empatica. (n.d.). Retrieved from https://www.empatica.com/research/e4/

[20]. Pakarinen, T., Pietilä, J., & Nieminen, H. (2019, July). Prediction of self-perceived stress and arousal based on electrodermal activity. In 2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC) (pp. 2191-2195). IEEE.

[21]. What EMBRACE2 Monitors – Empatica Support. (n.d.). Retrieved from https://support.empatica.com/hc/en-us/articles/360028338472-What-Embrace2-monitors

[22]. Umair, M., Chalabianloo, N., Sas, C., & Ersoy, C. (2021). HRV and stress: A mixed-methods approach for comparison of wearable heart rate sensors for biofeedback. IEEE Access, 9, 14005-14024.

[23]. Skaramagkas, V., Ktistakis, E., Manousos, D., Tachos, N. S., Kazantzaki, E., Tripoliti, E. E., ... & Tsiknakis, M. (2021, October). Cognitive workload level estimation based on eye tracking: A machine learning approach. In 2021 IEEE 21st International Conference on Bioinformatics and Bioengineering (BIBE) (pp. 1-5). IEEE.

[24]. Gao, J., Liu, S., Feng, Q., Zhang, X., Jiang, M., Wang, L., ... & Zhang, Q. (2019). Subjective and objective quantification of the effect of distraction on physician's workload and performance during simulated laparoscopic surgery. Medical Science Monitor: International Medical Journal of Experimental and Clinical Research, 25, 3127.

[25]. Makowski, D., Pham, T., Lau, Z. J., Brammer, J. C., Lespinasse, F., Pham, H., Schölzel, C., & Chen, S. A. (2021). NeuroKit2: A Python toolbox for neurophysiological signal processing. *Behavior Research Methods*, 53(4), 1689-1696. https://doi.org/10.3758/s13428-020-01516-y

[26]. Sanches, C. L., Augereau, O., & Kise, K. (2016, December). Manga content analysis using physiological signals. In Proceedings of the 1st international workshop on comics analysis, Processing and Understanding (pp. 1-6).

[27]. Coors, A., Breteler, M. M., & Ettinger, U. (2022). Processing speed, but not working memory or global cognition, is associated with pupil diameter during fixation. Psychophysiology, 59(11), e14089.

[28]. Tag, B., Vargo, A. W., Gupta, A., Chernyshov, G., Kunze, K., & Dingler, T. (2019, May). Continuous alertness assessments: Using EOG glasses to unobtrusively monitor fatigue levels In-The-Wild. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1-12).

[29]. Nourbakhsh, N., Wang, Y., & Chen, F. (2013). GSR and blink features for cognitive load classification. In Human-Computer Interaction–INTERACT 2013: 14th IFIP TC 13 International Conference, Cape Town, South Africa, September 2-6, 2013, Proceedings, Part I 14 (pp. 159-166). Springer Berlin Heidelberg.

[30]. Rahimi, A., & Recht, B. (2007). Random features for large-scale kernel machines. Advances in neural information processing systems, 20.

[31]. Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD explorations newsletter, 6(1), 20-29.

[32]. Fotouhi, S., Asadi, S., & Kattan, M. W. (2019). A comprehensive data level analysis for cancer diagnosis on imbalanced data. Journal of biomedical informatics, 90, 103089.

[33]. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

[34]. Romine, W. L., Schroeder, N. L., Graft, J., Yang, F., Sadeghi, R., Zabihimayvan, M., ... & Banerjee, T. (2020). Using machine learning to train a wearable device for measuring students' cognitive load during problem-solving activities based on electrodermal activity, body temperature, and heart rate: development of a cognitive load tracker for both personal and classroom use. Sensors, 20(17), 4833.