Applied Human Factors and Ergonomics (AHFE2025), Vol. 199, 2025, 700–713
https://doi.org/10.54941/ahfe1006883

A Systematic Review of Ground-Truth Labeling and Prediction for Cognitive Workload Adaptive Systems

Udit Kumar Das[1], Moajjem Chowdhury[1], Yunmei Liu[1], and David Kaber[2]

[1]Industrial and Systems Engineering, University of Louisville, Louisville, KY 40208, USA
[2]Mechanical, Industrial, and Manufacturing Engineering, Oregon State University, Corvallis, OR 97331, USA

ABSTRACT

Cognitive workload monitoring (or real-time inferencing) is crucial for the safe operation of complex human-machine systems, and motivates the development of adaptive automation technologies to dynamically assist operators and prevent both overload and disengagement situations. We systematically reviewed 75 recent studies (2015–2025) on machine learning-based cognitive workload monitoring and adaptive systems. The review focused on three key challenges: (1) ground-truth workload labeling; (2) predictive model generalization across users; and (3) adaptive automation/interface interventions. Approximately 28% of studies were found to rely on retrospective self-report workload scales for ground-truth labels, although some use objective task performance metrics or hybrid labeling approaches. Predictive models were observed to achieve high accuracy for the same individuals they were trained on (subject-dependent validation; mean ~85.6%), but performance dropped when tested on new users (subject-independent validation; mean ~80.3%). In general, the majority of studies present offline model development (for asynchronous classification of workload states) or conceptual system proposals; only 7 studies (9.3%) implemented and evaluated a real-time closed-loop workload-responsive system with human participants. These gaps highlight the need for standardized multimodal workload state labeling methods, cross-user modeling techniques, and empirical validation of closed-loop workload-adaptive systems in operational settings.

Keywords: Ground truth labeling, Cognitive workload, Predictive modeling, Adaptive interventions, Human-in-the-loop system

INTRODUCTION

Cognitive workload, the mental effort required to perform a task, is a critical factor influencing human performance and safety in complex systems. Both cognitive overload and underload (disengagement) can degrade operator effectiveness: overload may lead to errors, fatigue, or accidents, while underload can result in vigilance decrements, task boredom, and delayed responses (Young and Stanton, 2002; Wickens, 2008; Diarra et al., 2025).

Maintaining workload within an optimal range is therefore essential for high-risk domains, such as aviation, driving, air traffic control, and medicine (Grier et al., 2008). In response, researchers have increasingly focused on real-time mental workload (MWL) monitoring (i.e., inferencing on states during task execution) as a basis for adaptively automated systems that dynamically adjust forms of assistance or task allocations to keep operator workload at safe, efficient levels (Aricò et al., 2016). Adaptive automation aims to prevent performance decrements associated with extreme workload by maintaining task demands within a defined range to avoid both under- and over-load (Aricò et al., 2016). Achieving such closed-loop human-machine systems require reliable methods to continuously, objectively, and accurately assess cognitive workload and trigger task aids without disrupting human operator performance.

Traditionally, mental workload has been measured via retrospective self-report scales and post-task questionnaires. The NASA Task Load Index (NASA-TLX) is likely the most widely used example of a self-report index that asks operators to rate their perceived workload after completing a task (Hart and Staveland, 1988). While easy to administer, such subjective ratings introduce bias and cannot capture moment-to-moment fluctuations in cognitive load (Wu et al., 2025). Performance-based measures such as response times and error rates can also indicate workload, but they are often task-specific and may not generalize across contexts (Wickens, 1992). In recent years, advances in wearable sensors have enabled more objective, real-time workload assessment via physiological signals (Charles and Nixon, 2019). Our previous work has demonstrated biosignals such as brain activity (fNIRS), heart rate and heart rate variability, respiration, and eye metrics (pupil diameter, blink rate) to be highly sensitive to changes in cognitive demand and can be monitored continuously and unobtrusively (Liu et al., 2024a; Liu et al., 2024b; Wen et al., 2025; Grimaldi et al., 2024a; Grimaldi et al., 2024b; Nadri et al., 2024). A growing body of research applies machine learning to these multimodal data streams to predict an operator's cognitive workload state in real time (Grimaldi et al., 2024b). Recent survey articles reflect this shift toward objective and multimodal workload monitoring (Das Chakladar and Roy, 2024; Tao et al., 2019; Debie et al., 2021), noting that combining multiple modalities (e.g., EEG (electro-encephalography) with ECG (electro-cardiogram) and eye-tracking) can improve accuracy and robustness compared to single sensors.

Despite these advances, critical gaps (listed below) remain on the path to accurately predict cognitive workload and, furthermore, truly achieve adaptive workload management.

(1) In order to accurately predict workload, there is a need to define ground-truth workload levels. The human-machine systems field continues to rely heavily on subjective, post-hoc labeling of workload responses to train classification models, which are coarse and retrospective (Grimaldi et al., 2024b). Researchers have experimented with alternative labeling approaches, such as using task conditions or performance metrics as objective proxies, and developing hybrid schemes that fuse subjective

and objective indicators; however, no consensus has been achieved on a gold-standard labeling method (Tao et al. 2019).

(2) Most reported models have limited generalizability beyond their training conditions. Many machine-learning classifiers achieve high accuracy for the same individuals or tasks they were trained on, but performance often drops substantially when applied to new users or settings (Zhao et al., 2018; Boring et al., 2020; Debie et al., 2021; Sun and Li, 2025). High inter-individual variability in physiological responses means models tend to overfit to the idiosyncrasies of training data. Although some recent studies have shown improvements in cross-subject robustness by: (a) training on larger, more diverse datasets to reduce sampling bias; (b) validating transfer/domain-adaptation methods for cross-subject and cross-task EEG workload decoding; and (c) using deep subdomain adaptation to align class-conditional features with class-confusion loss (Sun and Li, 2025; Ding et al., 2023; Zhou et al., 2023; Luong et al., 2020; Wang et al. 2022). However, overall model predictive performance in novel contexts remains a concern (Debie et al., 2021).

(3) The integration of workload prediction into real-time adaptive system control is largely unachieved. Dozens of papers propose using workload estimates to trigger adaptive interventions (e.g., adjusting interface complexity or level of automation), yet only a small subset of research teams have actually implemented and evaluated such closed-loop systems with human operators (e.g., Lucchese et al., 2025). The deployment of adaptive workload-responsive systems in high-stakes domains is hindered by: (a) ambiguous MWL ground-truth labeling (Safari et al., 2024; Young et al., 2015); (b) artifact-prone sensing (via physiological measures) in operational settings (Aksu et al., 2024; Zhou et al., 2020; Schultze-Kraft et al., 2016; Hajra et al., 2020); (c) weak cross-subject/task generalization (Sun and Li, 2025; Zhou et al., 2023); (d) subjective labels and limited continuous ground-truth (Zhou et al., 2020; Liu Yisi et al., 2017); (e) safety risks from intrusive measurement approaches and automation false alarms (Zhou et al., 2020); and (f) a lab-to-field validity gap, including few demonstrations of research system use in actual applications (Kyle et al., 2025; Ding et al., 2023).

Published reviews have examined specific aspects of cognitive workload, such as particular sensors or data fusion methods, but none have focused on the intersection of ground-truth labeling, real-time predictive modeling, and adaptive (automation) interventions. To address this gap, we conducted a systematic review of recent research (2015–2025) on cognitive workload sensing/monitoring, machine learning-based state classification, and human-in-the-loop system adaptation. In particular, we investigated if (and how) contemporary systems effectively address these three challenges. We found 75 studies that attempted to establish reliable ground-truth MWL labels, generalize model-based predictions, and adapt automated assistance based on predictions. The overarching aim of these studies has been to synthesize current research trends and guide future efforts toward truly adaptive cognitive workload management in real-world applications.

METHODS

Search Method and Study Selection

We followed the PRISMA method to identify relevant studies (Page et al., 2021). The Web of Science database was searched for the period from January 1, 2015, to July 9, 2025. A keyword strategy, using the Boolean operators, combined terms for cognitive/mental workload with machine learning and terms capturing ground-truth/labeling, real-time monitoring, forecasting, or adaptive systems. In this review, we use the term prediction to refer to any form of predictive modeling (regression or classification) of MWL. To complement database results, backward and forward citation chasing was conducted on included studies.

This review focused on machine-learning systems that classify cognitive/mental workload states, based on sensor/response inputs and adapt to input patterns. Inclusion criteria required studies to: (a) address cognitive or mental workload; (b) apply machine learning for classification/prediction; and (c) describe labeling and/or real-time state assessment/adaptation. Studies also had to be peer-reviewed and presented in the English language. We excluded reviews, theses and editorials, non-English publications, studies absent of machine learning methods, and studies outside the scope of the review.

Initially, a total of 449 records were identified. After removing duplicate and irrelevant (out-of-scope) entries, 392 unique records remained for screening. Title and abstract screening excluded 315 records that did not meet the additional inclusion criteria, leaving 77 reports for full-text retrieval and review (with 10 being unobtainable). We assessed the remaining 67 full-text articles, of which 54 met all criteria. We then examined reference lists of the included papers, identifying 32 additional candidates, of which 21 were identified as eligible for further review. In total, 75 studies were included in the final review. (**Figure 1** shows the PRISMA flow diagram.)

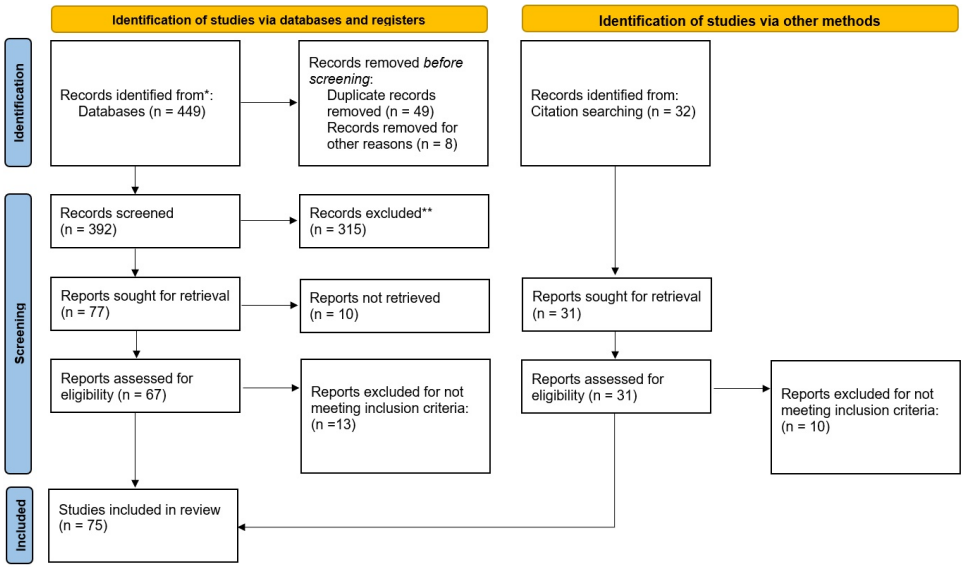

Figure 1: Prisma flow diagram.

We extracted key information from each study, including: application domain, ground-truth MWL labeling method(s) and tool and/or criteria, input modalities for prediction, ML model types/approaches, validation strategy, and real-time operation. We also recorded whether an adaptive intervention was implemented or proposed as part of a closed-loop system. The MWL labeling approach was categorized as subjective, objective, hybrid, or unsupervised/rule-based. The model validation type was classified as subject-dependent (data on the same individuals used for training and testing) or subject-independent (data on different participants used for training and testing). Workload processing modes were classified as "Post-hoc analysis" (data used solely for offline model development and validation) or "Real-time monitoring" (trained models applied for inference during live task execution). Regarding real-time adaptive interventions, the studies were classified as "Tested" (i.e., a closed-loop system), "Proposed" (a concept without implementation), or "None" (no adaptation considered).

RESULTS

Ground-Truth Labeling Methods

The reliability of workload predictions is contingent on the quality of ground-truth labels used during model training. **Table 1** summarizes the five primary labeling strategies reported across the corpus of research. Objective labels were most common and appeared in 41/75 articles (54.7%) with derivation based on task conditions, tiers of difficulty, or performance metrics. Subjective labels followed with 21/75 studies (28.0%) reporting use based on standardized self-reports. Rule-based/unsupervised labeling (6/75, 8.0%) and mixed subjective–objective labeling (4/75, 5.3%) were less frequent. Hybrid-fusion methods (3/75, 4.0%) were rare for labeling. This distribution suggests reliance on subjective labels remains widespread, despite retrospective bias and low temporal resolution. Objective and hybrid approaches are gaining traction but still lack standardization.

Table 1: Ground-truth labeling methods.

Ground-Truth Method	Labeling Tool	N	%
Objective	Task level (e.g., MATB/MATB-II), performance/SME rules, and occasional physiology thresholds. (MATB-II)	41	54.7%
Subjective	NASA-TLX/RTLX, SWAT, Paas, SAM; ratings binned to classes. (Human Factors NASA)	21	28.0%
Rule-based/ unsupervised	Fuzzy/if-then thresholds; clustering/SSL to derive states	6	8.0%
Subjective/ objective	Parallel subjective and objective labels used in separate analyses	4	5.3%
Hybrid	Fusion (weighted vote/rules) to form the final class.	3	4.0%

Classification of methods among categories was based on the "final type of label used for model training/testing"; self-reports used only for validation were not counted as "subjective labels".

Predictive Modeling

We found 57 workload classification studies that reported accuracy metrics, including 44 subject-dependent and 13 subject-independent. On average, subject-dependent models achieved 85.6% mean accuracy in classification (range 64.9–99.7%) relative to ground-truth labels, while subject-independent models averaged 80.3% accuracy (range 66.6–98.7%). Two patterns emerged from this part of the review:

(1) Subject-dependent analyses more often report very high performance, with 38% of studies reporting greater than 90% accuracy vs. only 8% of subject-independent studies doing so.
(2) Subject-independent studies have a heavier lower tail, with 54% of studies reporting less than 80% workload classification accuracy vs. only 27% for subject-dependent studies. These results reflect inter-subject variability: models trained and tested on the same persons can exploit stable, individual-specific response patterns; leave-one-subject-out exposes generalization limitations. For example, a subject-dependent study that achieved 95.29% accuracy, when classifying workload within the same level of task difficulty, produced only 72.2% accuracy for cross-level classification and 53.83% accuracy for cross-task classification (Zhao et al., 2018).

In addition, among subject-dependent studies, 32 utilized traditional supervised models, including SVM, Random Forest, and other tree-based models. In contrast, only 12 studies incorporated neural network-based models (CNN, RNN, and Graph NN). For subject-independent studies, seven employed traditional machine learning models, while six used neural networks. Among subject-independent studies, one investigation measured MWL using EEG measures for a sample of ten male volunteers and reported a classification accuracy of 98.66% when using an ANN (artificial neural network) trained on individualized alpha-frequency features (for 17 channels) and evaluated with leave-one-subject-out cross-validation (Samima and Sarma 2023). This was one of the more impressive results from our review, but the outcome benefits from a small, homogeneous cohort of participants and the simplicity of a 3-class n-back (working memory) task. In contrast, a high-density fNIRS study with 22 participants tested a multi-branch CNN-BiGRU-SLA model with a leave-one-participant-out validation approach and reported 0.9361 accuracy for binary working-memory load and 0.8948 accuracy for visual perceptual load. Multi-level classification accuracies dropped to 0.7994 for working-memory load and 0.7992 for visual perceptual load (Wang et al., 2022). These examples illustrate that subject-independent accuracy can be high when the sample is small, tasks are simple or binary, and response feature-engineering explicitly reduces cross-subject variability. Nonetheless, model performance generally decreases as the number of participants and the complexity of the task increase.

Seven workload classification studies reported "balance-aware" metrics rather than accuracy, including F1 scores, micro-F1 scores, the AUC metric, and model sensitivity. F1 scores ranged from 0.664 to 0.9998. For example, Agarwal et al., (2021) applied an LSTM (long-short-term memory model to

accelerometer/EDA/skin-temp/HR responses for an F1 score = 0.9998, but this result was for a single subject (i.e., subject-dependent modeling without generalizability). Similarly, Grimaldi et al., (2024a) used fNIRS response measures to classify workload responses with a weighted F1 score = 0.80 and a macro F1 score (on the test dataset) = 0.76. On the other hand, Luo et al., (2021) applied an HMM (Hidden-Markov Model) to eye-tracking gaze data, yielding an F1 score = 0.664 for subject-independent analysis. This result reflects the challenge of balanced multiclass/sequence labeling as a basis for workload classification. Studies applying regression analysis, eight studies reported MAE (mean absolute error) measures ranging from 0.1105 to 11.11, with the best result occurring for a dual-branch attention model using eye-tracking and PPG (pulse oximetry) measures in a subject-independent analysis (Wei et al., 2025). Regression study results also revealed MSE (mean square error) measures ranging from 0.17 to 1.487, and MAPE (mean absolute percentage error) measures from 0.015 to 0.36; with subject-independent setups generally producing less accurate classifications than subject-dependent.

Processing Mode and Adaptive Interventions

The majority of studies reviewed performed post-hoc analyses on pre-recorded data rather than functioning in real time. There were 64/75 studies (85.3%) that used offline processing (e.g., Liu et al., 2024a; Grimaldi et al., 2024a; Grimaldi et al., 2024b; Liu et al., 2024b; Vukovic et al., 2019; McKendrick et al., 2019); whereas, only 11/75 studies (14.7%) performed real-time monitoring (Wen et al., 2025; Jo et al., 2025; Wei et al., 2025; Yu et al., 2025; Yang et al., 2024; Sandoval et al., 2022; Luo et al., 2021; Planke et al., 2021; Luong et al., 2020; Lei et al., 2017; Aricò et al., 2016). In terms of adaptive interventions, 64/75 studies (85.3%) proposed a real-time intervention without implementation, and four studies (5.3%) did not propose real-time adaptation. Seven studies (9.3%) tested a real-time adaptive system in human–in-the-loop experiments (Wei et al., 2025; Wen et al., 2025; Jo et al., 2025; Lei et al., 2017; Luo et al., 2021; Aricò et al., 2016; Yang et al., 2024). **Table 2** summarizes this distribution. These numbers indicate that, despite frequent claims about adaptive or real-time technologies, the majority of published work remains at the analysis stage.

Table 2: Processing mode with adaptive intervention.

		Adaptive Intervention			
		Proposed	**Tested**	**None**	**Total**
Workload processing mode	Post-hoc analysis	60	0	4	64
	Real-time monitoring	4	7	0	11
	Total	64	7	4	75

Among the tested interventions, most studies explored secondary-task difficulty manipulations, changes in the level of primary task assistance, or automation state manipulations as the workload rose. Several studies reported lower workload and better task metrics under adaptation.

Regarding MWL labeling methods as a basis for model training and triggering adaptive interventions, McKendrick et al., (2019) formalized labeling by either dividing datasets based on task difficulty levels, using mixed-effects statistical models to account for differences in task difficulty, vs. Rasch models accounting for task difficulty as well as estimated individual capability. They showed that Rasch-based labels used with a Random Forest machine learning model resulted in superior AUC outcomes and supported cross-person/cross-task transfer. Wen et al., (2025) subsequently demonstrated that fNIRS-based indicators of cognitive workload input to an LLM-guided policy could be used to adaptively cue pilot attention in a VR cockpit. They tested eight licensed pilots with real-time adaptive visual, auditory, and textual cues. The system was labelled as the "AdaptiveCoPilot" and maintained pilots at an "optimal" level of workload for greater periods of time than baseline conditions, and improved task completion. In the driving domain, Lei et al., (2017) applied EEG-based workload state classification to adaptive task allocation. The system was capable of maintaining driver workload near "moderate" levels but with only modest performance gains, underscoring system feasibility more than effect size. Luo et al., (2021) applied an HMM to gaze-trajectory data plus eyes-on-road frequency and steering torque to scale haptic shared steering control in semi-autonomous driving. Human-in-the-loop simulator tests ($N = 24$) revealed lower driver workload, higher trust, better lane-keeping, and smaller control effort vs. a non-adaptive control condition.

Proposed systems commonly describe real-time workload monitors/-classifiers intended to trigger changes in the level of system automation, reallocate task loads (e.g., across operators or robots), or deliver adaptive feedback/training. However, these studies leave the closed-loop evaluation untested. The near-absence of tested interventions underscores the gap between predictive modeling research efforts and validated closed-loop human–machine system implementations.

DISCUSSION AND CONCLUSIONS

The findings of this review highlight both progress and persistent challenges in developing truly adaptive systems to respond to operator cognitive workload. Ground-truth labeling remains a foundational hurdle. Many studies still rely on retrospective self-report scales (e.g., NASA-TLX) as their only method of labeling operator workload responses, which are coarse, prone to recall bias, and lack the temporal resolution to capture rapid cognitive demand fluctuations during tasks. Machine learning model training on end-of-task subjective ratings typically only provides the capability for gross classifications (e.g., "easy" vs. "hard" conditions) rather than detecting dynamic and multi-level demand changes. This means that current models may be optimized to classify imprecise workload targets, providing weak capability to support automation interventions. Although over half of the studies we reviewed did adopt objective task-based labels, no consensus or

"gold standard" has emerged. Improving label fidelity through continuous or real-time rating methods, multimodal ground-truth measures (combining performance, physiological, and subjective indicators), or standardized benchmark tasks with known difficulty levels is essential to enhance classification/predictive model sensitivity and to allow for comparisons of models across studies.

Another key challenge to adaptive systems for cognitive workload is the generalizability of underlying models across users and contexts. Most machine learning models achieve high classification accuracy only in subject-dependent evaluations (trained and tested on the same individuals or task conditions). When applied to new participants or different task domains, performance often drops substantially, underscoring that many models may have been overfit to their training datasets. A multimodal anomaly-detection study achieved 95.3% accuracy in workload classification for the same level of task difficulty but only 53.8% when tested across task difficulties, highlighting the loss when conditions change (Zhao et al., 2018). Similarly, McKendrick et al., (2019) showed that Rasch labeling of cognitive workload responses, based on fNIRS data, combined with a Random Forest machine learning approach, yielded ROC-AUC values of ~0.91–0.92 for within-task cross-validation. However, these values dropped to ~0.81–0.82 for all held-out participants. Promisingly, a few studies have achieved cross-subject accuracies in the high 70-80% by using large, diverse datasets and domain adaptation methods, but such cases remain exceptions. To enable practical deployment, future research must prioritize robust validation on held-out users and conditions for testing. This includes normalizing physiological signals to individual baselines and developing feature extraction methods that account for inter-person variability. Emphasizing cross-user evaluation in model development will better reveal generalization limits and guide the creation of algorithms that maintain accuracy in highly variable real-world settings.

Perhaps the most significant research gap in this area is translating workload predictions to forms of real-time adaptive assistance. While the majority of papers conceptually propose adjusting various system functions based on workload fluctuations, only a small fraction have actually implemented and evaluated a closed-loop intervention with human users. These few proof-of-concept trials suggest that adaptive systems can indeed improve operator performance and regulate workload by, for example, modifying task difficulty, reducing interface complexity, or elevating the level of system automation when high workload is detected. However, these demonstrations have been limited to small-scale lab studies in specific domains, so it remains uncertain how well such approaches would generalize to other environments and underload responses. Several factors likely contribute to the paucity of tested interventions, including:

(1) integrating a workload model into a live control system is technically complex and resource-intensive;
(2) there are safety and usability concerns that an unreliable workload detector could trigger inappropriate actions (especially in high-stakes domains like aviation or healthcare); and

(3) evaluating closed-loop systems requires measuring not only prediction accuracy but also the impact of the adaptation on human performance, trust, or safety, which makes experimental design more challenging.

These barriers provide some explanation for why adaptive workload management remains largely conceptual in the scientific current literature.

Taken together, machine learning techniques for workload estimation have advanced to provide reliable real-time metrics, but their practical impact will remain limited unless the above three challenges can be addressed in parallel. The field of human-machine systems design and engineering needs to move beyond algorithm development toward system-level experimentation. Priorities should include establishing higher-fidelity, standardized MWL labeling methods, design of models explicitly for cross-user robustness, and conducting rigorous human-in-the-loop studies that test adaptive interventions in realistic tasks. Only by simultaneously improving label quality, model generalization, and closed-loop validation can cognitive workload–adaptive systems fulfill their promise of enhancing performance and safety in complex human–machine environments.

LIMITATION

One of the limitations of our study is that we only searched the Web of Science database and employed backward and forward citation chasing. Furthermore, coding ground-truth labeling methods occasionally require interpretation, and the clarity of the authors' reporting influenced our results. This might have introduced bias.

ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation under Grant Nos. CMMI-2535920 and CMMI-2535921. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

Agarwal, Ankita, Josephine Graft, Noah Schroeder, and William Romine. 2021. "Sensor-Based Prediction of Mental Effort during Learning from Physiological Data: A Longitudinal Case Study." Signals 2 (4): 886–901. https://doi.org/10.3390/signals2040051.

Aksu, Şeniz Harputlu, Erman Çakıt, and Metin Dağdeviren. 2024. "Mental Workload Assessment Using Machine Learning Techniques Based on EEG and Eye Tracking Data." Applied Sciences 14 (6): 2282. https://doi.org/10.3390/app14062282.

Aricò, Pietro, Gianluca Borghini, Gianluca Di Flumeri, et al., 2016. "Adaptive Automation Triggered by EEG-Based Mental Workload Index: A Passive Brain-Computer Interface Application in Realistic Air Traffic Control Environment." Frontiers in Human Neuroscience 10 (October): 539. https://doi.org/10.3389/fnhum.2016.00539.

Boring, Matthew J, Karl Ridgeway, Michael Shvartsman, and Tanya R Jonker. 2020. "Continuous Decoding of Cognitive Load from Electroencephalography Reveals Task-General and Task-Specific Correlates." Journal of Neural Engineering 17 (5): 056016. https://doi.org/10.1088/1741–2552/abb9bc.

Charles, Rebecca L., and Jim Nixon. 2019. "Measuring Mental Workload Using Physiological Measures: A Systematic Review." Applied Ergonomics 74 (January): 221–32. https://doi.org/10.1016/j.apergo.2018.08.028.

Das Chakladar, Debashis, and Partha Pratim Roy. 2024. "Cognitive Workload Estimation Using Physiological Measures: A Review." Cognitive Neurodynamics 18 (4): 1445–65. https://doi.org/10.1007/s11571–023-10051–3.

Debie, Essam, Raul Fernandez Rojas, Justin Fidock, et al., 2021. "Multimodal Fusion for Objective Assessment of Cognitive Workload: A Review." IEEE Transactions on Cybernetics 51 (3): 1542–55. https://doi.org/10.1109/TCYB.2019.2939399.

Diarra, Moussa, Jean Theurel, and Benjamin Paty. 2025. "Systematic Review of Neurophysiological Assessment Techniques and Metrics for Mental Workload Evaluation in Real-World Settings." Frontiers in Neuroergonomics 6 (April). https://doi.org/10.3389/fnrgo.2025.1584736.

Ding, Li, Jack Terwilliger, Aishni Parab, et al., 2023. "CLERA: A Unified Model for Joint Cognitive Load and Eye Region Analysis in the Wild." ACM Transactions on Computer-Human Interaction 30 (6): 1–23. https://doi.org/10.1145/3603622.

Grier, R., Wickens, C., Kaber, D., Strayer, D., Boehm-Davis, D., Trafton, J. G., & St. John, M. (2008). The Red-Line of Workload: Theory, Research, and Design. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *52*(18), 1204–1208. https://doi.org/10.1177/154193120805201811 (Original work published 2008).

Grimaldi, Nicolas, Yunmei Liu, Ryan McKendrick, Jaime Ruiz, and David Kaber. 2024a. "Deep Learning Forecast of Cognitive Workload Using fNIRS Data." 2024 IEEE 4th International Conference on Human-Machine Systems (ICHMS), May 15, 1–6. https://doi.org/10.1109/ICHMS59971.2024.10555701.

Grimaldi, Nicolas, David Kaber, Ryan McKendrick, and Yunmei Liu. 2024b. "Deep Learning Forecast of Perceptual Load Using fNIRS Data." Paper presented at 2024 AHFE International Conference on Human Factors in Design, Engineering, and Computing (AHFE 2024 Hawaii Edition). https://doi.org/10.54941/ahfe1005563.

Grimaldi, N., Kaber, D., Chen, Y., McKendrick, R., & Liu, Y. (2025, May). Deep Learning Forecast of Attention Using fNIRS Data. In *2025 IEEE 5th International Conference on Human-Machine Systems (ICHMS)* (pp.65–69). IEEE.

Hajra, Sujoy Ghosh, Pengcheng Xi, and Andrew Law. 2020. "A Comparison of ECG and EEG Metrics for In-Flight Monitoring of Helicopter Pilot Workload." 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), October 11, 4012–19. https://doi.org/10.1109/SMC42975.2020.9283499.

Hart, Sandra G., and Lowell E. Staveland. 1988. "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research." In Advances in Psychology, edited by Peter A. Hancock and Najmedin Meshkati, vol. 52. Human Mental Workload. North-Holland. https://doi.org/10.1016/S0166–4115(08)62386–9.

Jo, Wonse, Ruiqi Wang, Baijian Yang, Daniel Foti, Mo Rastgaar, and Byung-Cheol Min. 2025. "Cognitive Load-Based Affective Workload Allocation for Multihuman Multirobot Teams." IEEE Transactions on Human-Machine Systems 55 (1): 23–36. https://doi.org/10.1109/THMS.2024.3509223.

Kyle, Ainsley R., Brock Rouser, Ryan C. Paul, and Katherina A. Jurewicz. 2025. "Quantifying Pilot Performance and Mental Workload in Modern Aviation Systems: A Scoping Literature Review." Aerospace 12 (7): 626. https://doi.org/10.3390/aerospace12070626.

Lei, S., Takashi Toriizuka, and Mattias Roetting. 2017. "Driver Adaptive Task Allocation: A Field Driving Study:" Le Travail Humain Vol. 80 (1): 93–112. https://doi.org/10.3917/th.801.0093.

Liu, Y., Berman, J., Dodson, A., Park, J., Zahabi, M., Huang, H., ... & Kaber, D. B., 2024a. Human-centered evaluation of emg-based upper-limb prosthetic control modes. IEEE Transactions on Human-Machine Systems, 54(3), 271–281.

Liu Yisi, Trapsilawati Fitri, Hou Xiyuan, et al., 2017. "EEG-Based Mental Workload Recognition in Human Factors Evaluation of Future Air Traffic Control Systems." In Advances in Transdisciplinary Engineering. IOS Press. https://doi.org/10.3233/978-1-61499-779-5-357.

Liu, Yunmei, Nicolas S. Grimaldi, Niosh Basnet, et al., 2024b. "Classifying Cognitive Workload Using Machine Learning Techniques and Non-Intrusive Wearable Devices." 2024 IEEE 4th International Conference on Human-Machine Systems (ICHMS), May 15, 1–6. https://doi.org/10.1109/ICHMS59971.2024.10555690.

Lucchese, Andrea, Antonio Padovano, and Francesco Facchini. 2025. "Comprehensive Systematic Literature Review on Cognitive Workload: Trends on Methods, Technologies and Case Studies." IET Collaborative Intelligent Manufacturing 7 (1): e70025. https://doi.org/10.1049/cim2.70025.

Luo, Ruikun, Yifan Weng, Yifan Wang, et al., 2021. "A Workload Adaptive Haptic Shared Control Scheme for Semi-Autonomous Driving." Accident Analysis & Prevention 152 (March): 105968. https://doi.org/10.1016/j.aap.2020.105968.

Luong, Tiffany, Nicolas Martin, Anais Raison, Ferran Argelaguet, Jean-Marc Diverrez, and Anatole Lecuyer. 2020. "Towards Real-Time Recognition of Users Mental Workload Using Integrated Physiological Sensors Into a VR HMD." 2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), November, 425–37. https://doi.org/10.1109/ISMAR50242.2020.00068.

McKendrick, Ryan, Bradley Feest, Amanda Harwood, and Brian Falcone. 2019. "Theories and Methods for Labeling Cognitive Workload: Classification and Transfer Learning." Frontiers in Human Neuroscience 13 (September): 295. https://doi.org/10.3389/fnhum.2019.00295.

Nadri, Chihab, Yunmei Liu, Maryam Zahabi, et al., 2024. "Analysis of Pre-Flight and Monitoring Tasks Using Cognitive Performance Modeling." Human Factors in Design, Engineering, and Computing 159 (159). https://doi.org/10.54941/ahfe1005693.

Page, Matthew J., Joanne E. McKenzie, Patrick M. Bossuyt, et al., 2021. "The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews." Research Methods & Reporting. BMJ 372 (March): n71. https://doi.org/10.1136/bmj.n71.

Planke, Lars J., Alessandro Gardi, Roberto Sabatini, Trevor Kistan, and Neta Ezer. 2021. "Online Multimodal Inference of Mental Workload for Cognitive Human Machine Systems." Computers 10 (6): 81. https://doi.org/10.3390/computers10060081.

Safari, MohammadReza, Reza Shalbaf, Sara Bagherzadeh, and Ahmad Shalbaf. 2024. "Classification of Mental Workload Using Brain Connectivity and Machine Learning on Electroencephalogram Data." Scientific Reports 14 (1): 9153. https://doi.org/10.1038/s41598–024-59652-w.

Samima, Shabnam, and Monalisa Sarma. 2023. "Mental Workload Level Assessment Based on Compounded Hysteresis Effect." Cognitive Neurodynamics 17 (2): 357–72. https://doi.org/10.1007/s11571–022-09830–1.

Sandoval, Catherine, Melissa N. Stolar, Simon G. Hosking, Dawei Jia, and Margaret Lech. 2022. "Real-Time Team Performance and Workload Prediction From Voice Communications." IEEE Access 10: 78484–92. https://doi.org/10.1109/ACCESS.2022.3193694.

Schultze-Kraft, Matthias, Sven Dähne, Manfred Gugler, Gabriel Curio, and Benjamin Blankertz. 2016. "Unsupervised Classification of Operator Workload from Brain Signals." Journal of Neural Engineering 13 (3): 036008. https://doi.org/10.1088/1741–2560/13/3/036008.

Sun, Wu, and Junhua Li. 2025. "AdaptEEG: A Deep Subdomain Adaptation Network With Class Confusion Loss for Cross-Subject Mental Workload Classification." IEEE Journal of Biomedical and Health Informatics 29 (3): 1940–49. https://doi.org/10.1109/JBHI.2024.3513038.

Tao, Da, Haibo Tan, Hailiang Wang, Xu Zhang, Xingda Qu, and Tingru Zhang. 2019. "A Systematic Review of Physiological Measures of Mental Workload." International Journal of Environmental Research and Public Health 16 (15): 2716. https://doi.org/10.3390/ijerph16152716.

Vukovic, Maria, Vidhyasaharan Sethu, Jessica Parker, Lawrence Cavedon, Margaret Lech, and John Thangarajah. 2019. "Estimating Cognitive Load from Speech Gathered in a Complex Real-Life Training Exercise." International Journal of Human-Computer Studies 124 (April): 116–33. https://doi.org/10.1016/j.ijhcs.2018.12.003.

Wang, Jiyang, Trevor Grant, Senem Velipasalar, Baocheng Geng, and Leanne Hirshfield. 2022. "Taking a Deeper Look at the Brain: Predicting Visual Perceptual and Working Memory Load From High-Density fNIRS Data." IEEE Journal of Biomedical and Health Informatics 26 (5): 2308–19. https://doi.org/10.1109/JBHI.2021.3133871.

Wei, Jishang, Erika Siegel, Prahalathan Sundaramoorthy, et al., 2025. "Cognitive Load Inference Using Physiological Markers in Virtual Reality." 2025 IEEE Conference Virtual Reality and 3D User Interfaces (VR), March 8, 759–69. https://doi.org/10.1109/VR59515.2025.00098.

Wen, Shaoyue, Michael Middleton, Songming Ping, et al., 2025. "AdaptiveCoPilot: Design and Testing of a NeuroAdaptive LLM Cockpit Guidance System in Both Novice and Expert Pilots." 2025 IEEE Conference Virtual Reality and 3D User Interfaces (VR), March 8, 656–66. https://doi.org/10.1109/VR59515.2025.00088.

Wickens, C. D. (1992). *Engineering psychology and human performance* (2nd ed.). HarperCollins Publishers. Wickens, Christopher D. 2008. "Multiple Resources and Mental Workload." Human Factors 50 (3): 449–55. https://doi.org/10.1518/001872008X288394.

Wu, Yun, Yile Zhu, and Bin Zheng. 2025. "Enhancing Surgical Training through Cognitive Load Assessment." Laparoscopic, Endoscopic and Robotic Surgery, ahead of print, June 25. https://doi.org/10.1016/j.lers.2025.06.001.

Yang, Jing, Juan Antonio Barragan, Jason Michael Farrow, Chandru P. Sundaram, Juan P. Wachs, and Denny Yu. 2024. "An Adaptive Human-Robotic Interaction Architecture for Augmenting Surgery Performance Using Real-Time Workload Sensing—Demonstration of a Semi-Autonomous Suction Tool." Human Factors: The Journal of the Human Factors and Ergonomics Society 66 (4): 1081–102. https://doi.org/10.1177/00187208221129940.

Young, Mark S., and Neville A. Stanton. 2002. "Attention and Automation: New Perspectives on Mental Underload and Performance." Theoretical Issues in Ergonomics Science 3 (2): 178–94. https://doi.org/10.1080/14639220210123789.

Young, Mark S., Karel A. Brookhuis, Christopher D. Wickens, and Peter A. Hancock. 2015. "State of Science: Mental Workload in Ergonomics." Ergonomics 58 (1): 1–17. https://doi.org/10.1080/00140139.2014.956151.

Yu, Xiaoqing, Chun-Hsien Chen, and Haohan Yang. 2025. "Cognitive Workload Quantification for Air Traffic Controllers: An Ensemble Semi-Supervised Learning Approach." Advanced Engineering Informatics 64 (March): 103065. https://doi.org/10.1016/j.aei.2024.103065.

Zhao, Guozhen, Yong-Jin Liu, and Yuanchun Shi. 2018. "Real-Time Assessment of the Cross-Task Mental Workload Using Physiological Measures During Anomaly Detection." IEEE Transactions on Human-Machine Systems 48 (2): 149–60. https://doi.org/10.1109/THMS.2018.2803025.

Zhou, Tian, Jackie S. Cha, Glebys Gonzalez, Juan P. Wachs, Chandru P. Sundaram, and Denny Yu. 2020. "Multimodal Physiological Signals for Workload Prediction in Robot-Assisted Surgery." ACM Transactions on Human-Robot Interaction 9 (2): 1–26. https://doi.org/10.1145/3368589.

Zhou, Yueying, Pengpai Wang, Peiliang Gong, et al., 2023. "Cross-Subject Cognitive Workload Recognition Based on EEG and Deep Domain Adaptation." IEEE Transactions on Instrumentation and Measurement 72: 1–12. https://doi.org/10.1109/TIM.2023.3276515.