Neuroergonomics & Cognitive Engineering, Vol. 207, 2026, 66–76
https://doi.org/10.54941/ahfe1007396

Multimodal Assessment of Pilot Cognitive Workload Using ECG and Eye-Tracking Features in Simulated Flight Tasks

Yixuan Guan[1], Longzhu Han[2], Pengyan Zhou[2], and Jiayi Bao[2]

[1]Hangzhou International Innovation Institute, Beihang University, Hangzhou, 311115, China

[2]School of Aeronautic Science and Engineering, Beihang University, Beijing, 100191, China

ABSTRACT

Pilot cognitive workload is a critical determinant of task performance, decision-making quality, and overall flight safety, rendering its objective and accurate assessment essential. However, current cognitive load assessment methods rely mainly on electroencephalography (EEG) signals that are easily disturbed, which limits their real-world use. Therefore, this study focuses on assessing pilot cognitive workload in realistic application scenarios, and proposed a multimodal assessment method of pilot cognitive workload base on electrocardiography (ECG) and eye-tracking signals, which their have high reliability and applicability. Firstly, flight tasks with graduated difficulty were designed in a simulated environment to induce three distinct levels of cognitive workload, and at the same time, the pilots are required to complete corresponding Overall Workload Scale (OWL) and the NASA Task Load Index (NASA-TLX) scale in different flight tasks. In the objective assessment, the multimodal signals of the pilot were collected based on ECG and eye-tracking signals for feature extraction. Then, effective feature indicators were selected using ANOVA, and redundant features were removed through Spearman correlation analysis. Finally, a multimodal pilot cognitive workload assessment model based on ECG and eye-tracking signals was built using a stacking ensemble learning framework with decision-level fusion. Results demonstrated the effectiveness of the workload task paradigm. Subjective ratings increased progressively with task difficulty, as confirmed by both NASA-TLX and OWL. Physiological analysis revealed distinct trends under increased workload. For ECG metrics, heart rate rose significantly, while heart rate variability indices—specifically Mean NN, RMSSD, LF, HF, TP, and Lempel-Ziv complexity—demonstrated significant decreases. Regarding ocular metrics, fixation duration, blink interval, and pupil diameter increased significantly, whereas saccade duration and blink frequency decreased. Following feature refinement via Spearman correlation analysis to remove redundancy (coefficient > 0.8), eleven key features were retained: six ECG features (Heart Rate, RMSSD, LF, HF, ApEn, and Lempel-Ziv complexity) and five eye-tracking features (Fixation Duration, Saccade Duration, Blink Count, Blink Interval, and Pupil Diameter). Based on these refined features, the multimodal assessment model, built with a stacking ensemble framework using decision-level fusion and employing Random Forest and K-Nearest Neighbors as base classifiers for ECG and eye-tracking data, achieved an accuracy of 0.959. Collectively, these findings substantiate the validity and sensitivity of fusing ECG and ocular metrics, providing critical data and technical support for the advancement of pilot physiological monitoring and early-warning systems.

Received February 5, 2026; Revised April 4, 2026; Accepted April 22, 2026; Available online July 20, 2026

Keywords: Cognitive workload assessment, Human factors in aviation, Electrocardiogram (ECG), Eye tracking, Multimodal fusion model

INTRODUCTION

Cognitive workload refers to the mental resources required to meet task demands during perception, decision making, and control (Kosch et al., 2023). In flight operations and training, cognitive workload is critical because it directly constrains information processing and control performance (K et al., 2020) and serves as a key state input for human-machine interface design and human-machine function allocation (Lim et al., 2018; Buerkle et al., 2022; Pei et al., 2024).

Previous studies have shown that cognitive workload is associated with changes in performance, situational awareness, and decision quality during flight tasks (Nadri et al., 2024). Manipulations of task structure and control demands can induce variations in cognitive workload (Mohanavelu et al., 2022). However, in structured flight tasks, the reliability and sensitivity with which cognitive workload differences induced by task manipulations can be captured across task conditions remain insufficiently characterized, particularly when task difficulty levels are closely spaced and cognitive demands change dynamically (K et al., 2020). Under such conditions, task performance is often maintained, which further complicates the detection of subtle workload variations.

In this context, existing approaches to cognitive workload assessment in aviation mainly rely on subjective rating scales, task performance metrics, and physiological signals. Subjective measures provide direct insights into perceived workload but suffer from recall bias and limited temporal resolution (Nikolai von et al., 2022). Performance-based indicators reflect task outcomes but are often insensitive to internal cognitive state changes when performance is maintained (Nadri et al., 2024). As a result, physiological measures have been increasingly adopted because they enable continuous and objective observation of internal cognitive states during task execution (Boere et al., 2024; Das Chakladar and Roy, 2024). Existing studies in aviation and related domains have investigated workload-related changes using a range of physiological signals, including electroencephalograph (EEG), eye movement, and electrocardiography (ECG). EEG-based features, such as spectral power distributions and connectivity patterns, are sensitive to changes in cognitive workload (Taheri Gorji et al., 2023; Naseri et al., 2025). In contrast, eye-tracking measures provide information about how attention is allocated and how visual scanning strategies change as task demands increase (Efetürk et al., 2022; Li et al., 2025). Workload-related changes in autonomic regulation have been examined using ECG, which reflects physiological responses associated with mental effort and stress (Bipro et al., 2023; Guo and Wu, 2025). Together, these findings suggest that physiological signals can reveal workload variations that may not be evident from performance measures alone, particularly when task performance remains stable.

However, current research predominantly relies on single-modality physiological signals, and multi-physiological workload assessment approaches still largely depend on EEG as the primary information source (Aygun et al., 2022; Chen et al., 2025; Liu et al., 2025). Although EEG is sensitive to cognitive state changes, it is vulnerable to motion artifacts and environmental interference (Hossain et al., 2022), which limits its robustness in operational settings. These constraints pose challenges for practical deployment in real flight environments. Therefore, considering feasibility and applicability, the present study adopts a multi-physiological approach based on eye-tracking and ECG to assess cognitive workload in structured and dynamic flight tasks.

EXPERIMENTAL DESIGN

Material and Procedure

Flight simulation was conducted using the DCS World software (Eagle Dynamic SA) with the Su-27 fighter aircraft model. Three flight tasks of low, medium, and high difficulty were designed to elicit corresponding levels of cognitive workload.

Low difficulty: A 10 min daytime visual cruise flight under clear, windless conditions at 650 kt and 1,500 ft, passing four predefined waypoints.

Medium difficulty: A 10 min standard traffic pattern under daytime visual meteorological conditions at 270 kt and 1,500 ft (Liu et al., 2023).

High difficulty: A 10 min standard traffic pattern under instrument flight conditions in cloudy, windy weather, with a 22 kt crosswind and 1,500 ft visibility.

The experiment was structured into three stages, each containing low, medium, and high difficulty flight tasks. To minimize order effects, the three difficulty conditions within each stage were presented in a randomized sequence. Each difficulty level was completed once per stage, and participants filled out the Over Workload Scale (OWL) immediately after each task to report their perceived workload. After completing all stages, participants also completed the NASA-TLX scale to provide a comprehensive workload evaluation for each difficulty condition. The experimental procedure is shown in Figure 1.

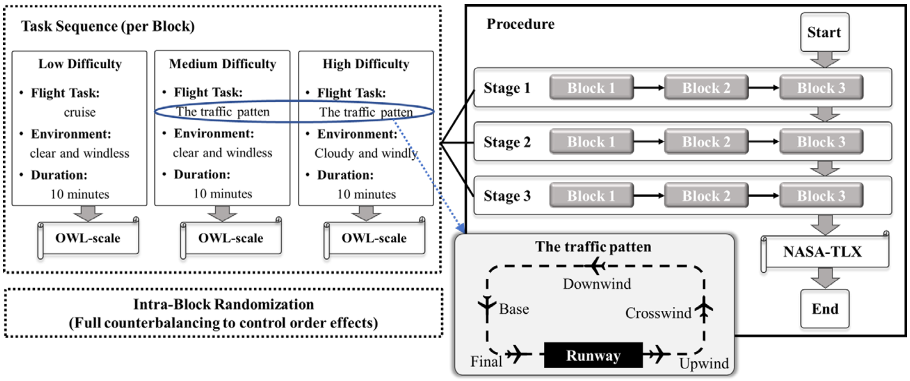

Figure 1: Procedure of the experiment.

During all flight tasks, physiological data were continuously recorded to assess cognitive workload. Eye-tracking data were collected using Tobii Pro Glasses 3 at a sampling rate of 100 Hz. The eye-tracking device was worn by participants throughout task execution to capture gaze behaviour during flight operations. ECG signals were recorded using a BIOPAC MP160 system (BIOPAC Systems, Inc., Goleta, CA, USA) with an RSPEC-R amplifier module at a sampling rate of 500Hz. Both physiological data streams were recorded continuously across all experimental stages and were temporally aligned with the flight simulation timeline to ensure correspondence between task execution and physiological responses.

Participants

A total of 30 male pilot trainees participated in this study, with an average age of 23.0 years (SD = 2.26). All participants had normal or corrected-to-normal vision, normal hearing, and no history of psychiatric or neurological disorders. They were instructed to avoid alcohol and caffeine before the experiment and to maintain at least eight hours of sleep.

Before data collection, all participants received training and signed informed consent forms to ensure familiarity with the flight simulator and a full understanding of the experimental tasks. The experiment was conducted at Beihang University and was approved by the Biological and Medical Ethics Committee of Beihang University.

EXPERIMENTAL RESULTS

Subjective Data

For subjective data, reliability of the Overall Workload Scale (OWL) was verified using the intraclass correlation coefficient (ICC). ICC values across all difficulty levels exceeded 0.5 (low: 0.596; medium: 0.501; high: 0.730), confirming satisfactory repeatability. Averaged OWL scale scores, which followed a normal distribution (Shapiro–Wilk test), were analysed via one-way ANOVA and Tukey's post hoc tests. Results revealed significant differences across the three task levels ($p < 0.05$), with distinct pairwise differences between all conditions (see Figure 2).

NASA-TLX scale scores were calculated using a weighted pairwise comparison of its six dimensions. Following a Shapiro–Wilk test, a one-way ANOVA and Tukey's post hoc tests revealed highly significant differences in overall workload across the three difficulty levels ($p < 0.0001$) , as shown in Figure 3. Furthermore, because cognitive workload is conceptually tied to the "mental demand" dimension, this sub-scale was analysed separately. As mental demand scores were non-normally distributed, non-parametric tests were employed, revealing highly significant variations ($p < 0.0001$) that were more pronounced than other dimensions. These findings indicate that while both overall workload and mental demand scaled with task difficulty, mental demand served as a more sensitive indicator of the cognitive load shifts induced by the experimental design. These results demonstrate that the experimental tasks effectively induced the intended cognitive workload levels, validating the sensitivity and rationality of the task design.

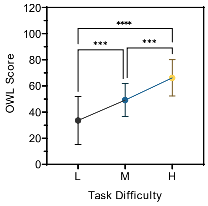

Figure 2: Relationship between OWL scale scores and task difficulty.

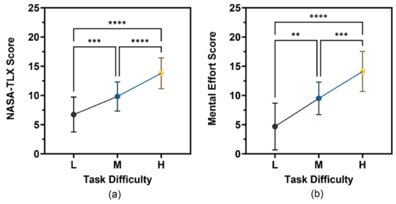

Figure 3: Overall NASA-TLX scale scores and mental demand scores across different task difficulty levels.

ECG Data

To enhance signal stability and R-wave detection accuracy, raw ECG data were processed using a 0.5 Hz high-pass filter to eliminate respiratory and motion-induced baseline drift, followed by a 50 Hz notch filter to remove power-line interference. A 20 Hz low-pass filter was then applied to suppress high-frequency noise. Post-preprocessing, 11 ECG-derived features were extracted using a 1-minute sliding window with a 30 s step size, including time-domain (HR, MeanNN, SDNN, RMSSD)(Chen et al., 2025), frequency-domain (LF, HF, TP, LF/HF), and nonlinear metrics (ApEn, FuzzyEn, LZC) (Germán-Salló, 2019).

One-way ANOVA was performed to analyze the ECG-derived indices across low, medium, and high workload conditions, and the results are shown in Figure 4 below. As could be seen from Figure 4, increasing task difficulty induced significant modulation of the autonomic nervous system.

Specifically, under high-load conditions, the significant increase in HR ($p < 0.05$) alongside the reduction in MeanNN ($p < 0.05$) and RMSSD indicated heightened sympathetic activation and attenuated parasympathetic regulation. In the frequency domain, a general decline in LF, HF, and TP, combined with an elevated LF/HF ratio, reflected a suppression of total autonomic activity and a distinct shift toward sympathetic dominance. Furthermore, the increase in ApEn suggested greater short-term irregularity in heart rate sequences under stress, whereas the decrease in other nonlinear metrics pointed to a reduction in overall signal complexity.

To ensure the quality of the final feature set, metrics were filtered based on statistical significance and redundancy. Regarding the ECG data, non-significant features were excluded based on the preceding significance analysis. A Spearman correlation analysis was then performed on the remaining features to identify and eliminate redundancy, with a correlation coefficient threshold of 0.8. As illustrated in the Figure 5, six features were ultimately retained: HR, RMSSD, LF, HF, ApEn, and LZC.

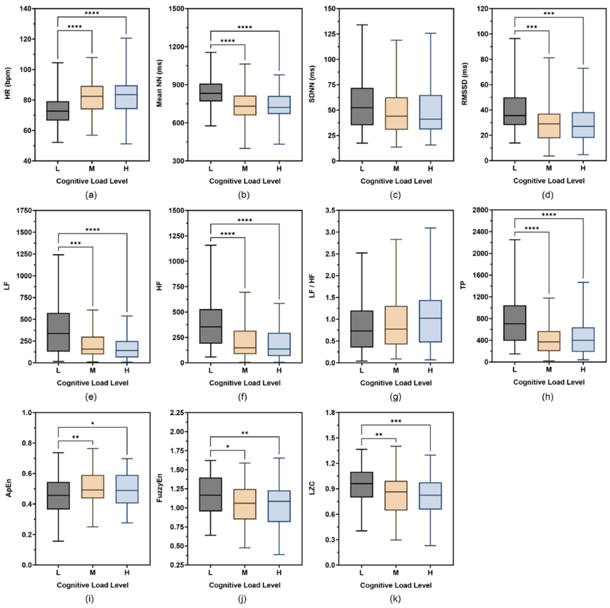

Figure 4: Differences in ECG indicators across cognitive workload levels.

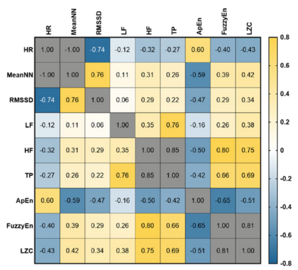

Figure 5: Results of correlation analysis of ECG indicators.

Eye-Tracking Data

Eye-tracking data were initially imported into the Tobii Pro Lab software, where valid segments were isolated based on the onset and completion times of the experimental tasks. Following standardized preprocessing, including filtering and smoothing to ensure data integrity, seven representative metrics were extracted using a 1-minute sliding window (see Table 1)(Chen et al., 2025).

To evaluate the sensitivity of these metrics, a one-way ANOVA was performed across the three conditions. The results are shown in Figure 6 below. As could be seen from Figure 6, all eye-tracking metrics exhibited significant differences across varying cognitive workload levels ($p < 0.05$). These findings demonstrate that the selected eye-tracking indices effectively characterize shifts in cognitive demand, validating their utility as objective measures of task-induced workload.

Building upon these significant results, a correlation analysis was subsequently conducted to address collinearity, as illustrated in the figure below. After removing redundant variables, five features were preserved: Fixation duration, Scanning duration, Blink frequency, Blink interval, and Pupil diameter.

Table 1: Eye-tracking metrics.

Metrics	Description
Fixation duration (s)	The cumulative duration of all fixation behaviours within a 1-minute window
Fixation frequency	The number of fixations within a 1-minute window
Scanning duration (s)	The cumulative duration of scanning within 1-minute
Scanning frequency	The frequency at which eye movement rapidly jumps from one fixation point to another within 1 minute
Blink frequency	The number of blinks within 1-minute
Blink interval (s)	The average time interval between two consecutive blinks
Pupil diameter (mm)	The average pupil size within 1-minute

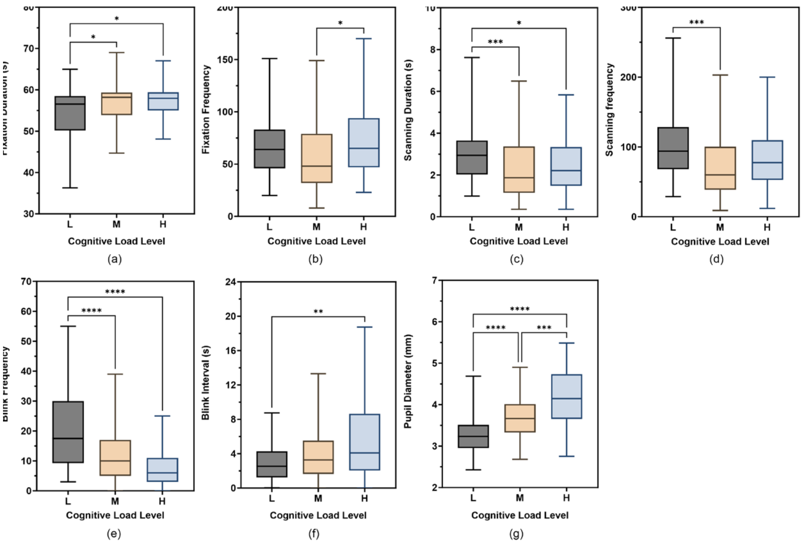

Figure 6: Differences in Eye-tracking metrics across cognitive workload levels.

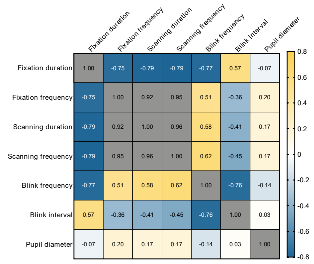

Figure 7: Results of correlation analysis of Eye-tracking metrics.

COGNITIVE LOAD ASSESSMENT MODEL

To avoid redundancy and noise propagation caused by direct feature concatenation, and to preserve the independence and complementarity of each modality, a Stacking-based multimodal fusion approach was employed for decision-level integration of eye-tracking and ECG data(Wang et al., 2025).

Initially, cognitive workload was modelled for individual modalities using five machine learning methods: logistic regression (LR), k-nearest neighbours (KNN), support vector machine (SVM), random forest (RF), and adaptive boosting (AdaBoost). Comprehensive evaluations indicated that RF achieved the highest standalone performance across both modalities, as shown in Table 2 and Table 3.

Table 2: Performance of ECG-based models for cognitive workload classification.

Algorithm	Precision	Accuracy	Recall	F1 Score
LR	0.378	0.408	0.363	0.316
KNN	0.707	0.708	0.705	0.706
SVM	0.473	0.499	0.465	0.456
RF	**0.881**	**0.889**	**0.879**	**0.881**
AdaBoost	0.803	0.803	0.802	0.802

Table 3: Performance of eye-tracking-based models for cognitive workload classification.

Algorithm	Precision	Accuracy	Recall	F1 Score
LR	0.511	0.496	0.497	0.483
KNN	0.667	0.664	0.660	0.659
SVM	0.609	0.604	0.602	0.602
RF	**0.710**	**0.707**	**0.705**	**0.705**
AdaBoost	0.567	0.560	0.560	0.560

However, to enhance model diversity within the Stacking framework, RF was selected for the ECG modality, while KNN was chosen for the eye-tracking modality. Despite RF's superior standalone accuracy for eye-tracking, integrating a distance-based learner (KNN) with a tree-based ensemble learner (RF) creates a heterogeneous ensemble. This configuration leverages complementary decision logics to capture distinct feature patterns, thereby reducing the correlation of prediction errors and minimizing generalization risk. For the meta-learner, Logistic Regression (LR) was employed to integrate the outputs from the base learners. This fusion approach yielded superior performance, achieving an accuracy of 0.9599, precision of 0.9598, recall of 0.9589, and an F1-score of 0.9593.

These results demonstrate that decision-level multimodal fusion can effectively exploit the complementary characteristics of eye-tracking and ECG signals, leading to more robust and reliable cognitive workload classification than single-modality models.

CONCLUSION

This study demonstrates that both electrocardiographic and ocular physiological signals provide effective representations of cognitive workload transitions during structured flight tasks. Increased workload was consistently associated with elevated heart rate and pronounced reductions in heart rate variability indices, including MeanNN, RMSSD, and frequency-domain parameters. In parallel, eye-tracking measures exhibited longer fixation durations, larger pupil diameters, and extended blink intervals, accompanied by reduced saccade duration and blink frequency. Together, these convergent physiological responses confirm the discriminative capability of both ECG and eye-tracking modalities for cognitive workload assessment.

Among the five evaluated machine learning methods, Random Forest (RF) achieved the highest classification accuracy for both ECG- and eye-tracking-based single-modality models. The ensemble tree architecture of RF proved particularly effective in capturing non-linear relationships and complex statistical patterns within physiological data streams, indicating its suitability for standalone cognitive workload modelling.

Building upon these single-modality results, the proposed Stacking-based decision-level fusion framework further improved classification robustness and generalization performance. By integrating heterogeneous base learners—RF for ECG and KNN for eye-tracking—with a Logistic Regression meta-learner, the fusion strategy leveraged complementary decision mechanisms while reducing error correlation and overfitting risk. As a result, the multimodal fusion approach achieved superior overall performance compared to individual modalities, highlighting its effectiveness for reliable cognitive workload recognition in dynamic and structured task environments.

ACKNOWLEDGMENT

Gratitude is extended to the Digital Aircraft Education and Research Platform of Beihang University for providing the flight simulator used in this study. Appreciation is also extended to all participants for their active participation and cooperation during the experiments.

REFERENCES

Aygun, A. Lyu, B. Nguyen, T. Haga, Z. Aeron, S. Scheutz, M. (2022) "Cognitive Workload Assessment via Eye Gaze and EEG in An Interactive Multi-modal Driving Task", Proceedings of the 2022 International Conference on Multimodal Interaction, ICMI '22. Association for Computing Machinery, New York, NY, USA, doi: 10.1145/3536221.3556610.

Bipro, S.S. Kabir, S and Motin, M.A. (2023) "Monitoring Stress Using Electrocardiogram Signal", 2023 Computing in Cardiology (CinC), Atlanta, GA, USA, doi: 10.22489/CinC.2023.204.

Boere, K. Anderson, F. Hecker, K.G. Krigolson, O.E. (2024). Measuring Cognitive Load in Multitasking using Mobile fNIRS, NeuroImage: Reports, Volume 4, No. 4, doi: 10.1016/j.ynirp.2024.100228.

Buerkle, A. Matharu, H. Al-Yacoub, A. Lohse, N. Bamber, T. Ferreira, P. (2022). An Adaptive Human Sensor Framework for Human–robot Collaboration, The International Journal of Advanced Manufacturing Technology, Volume 119, No. 1, pp. 1233–1248, doi: 10.1007/s00170-021-08299-2.

Chen, B. Yu, Y. Ding, L. (2025). Correlation Analysis of Physiological Parameters of Pilot Mental Workload Based on Flight Simulation, Biomedical Signal Processing and Control, Volume 109, pp. 108024, doi: 10.1016/j.bspc.2025.108024.

Das Chakladar, D. Roy, P.P. (2024). Cognitive Workload Estimation Using Physiological Measures: A Review, Cognitive Neurodynamics, Volume 18, No. 4, pp. 1445–1465, doi: 10.1007/s11571-023-10051-3.

Efetürk, Ö.A. Turgut, G. Dereshgi, H.A. Yılmaz, A. (2022). A Review of Visual Attention Research Using Eye-tracking Technologies, Journal of Smart Systems Research, Volume 3, No. 2, pp. 128–135.

Germán-Salló, Z. (2019) "HRV Analysis Using Complexity Indices", Presented at the 2019 E-Health and Bioengineering Conference (EHB), pp. 1–4, doi: 10.1109/EHB47216.2019.8969875.

Guo, C. Wu, J. Jing, Z. Zhan, X. Zhu, Z.H. (2025) "Investigation of the Relationship between Cognitive Load and Pilot ECG Characteristics", Proceedings of the International Conference on Aerospace System Science and Engineering 2024, Springer Nature, Singapore, pp. 425–433, doi: 10.1007/978-981-96-2440-9_42.

Hossain, Md.S. Reaz, M.B.I. Chowdhury, M.E.H. Ali, S.H.M. Bakar, A.A.A. Kiranyaz, S. Khandakar, A. Alhatou, M. Habib, R. (2022). Motion Artifacts Correction from EEG and fNIRS Signals Using Novel Multiresolution Analysis, IEEE Access, Volume 10, pp. 29760–29777, doi: 10.1109/ACCESS.2022.3159155.

K, M. S, P. K, A. D, R. Chinnadurai, V. S, V. K, R. Jayaraman, S. (2020). Dynamic Cognitive Workload Assessment for Fighter Pilots in Simulated Fighter Aircraft Environment Using EEG, Biomedical Signal Processing and Control Volume 61, p. 102018, doi: 10.1016/j.bspc.2020.102018.

Kosch, T. Karolus, J. Zagermann, J. Reiterer, H. Schmidt, A. Woźniak, P.W. (2023). A Survey on Measuring Cognitive Workload in Human-computer Interaction, ACM Comput. Surv., Volume 55, No. 13s, pp. 283:1-283:39, doi: 10.1145/3582272.

Li, Y. Zheng, B. Grant, P. Li, X. (2025). Assessing Cognitive Load through Eye Metrics in In-motion vs. Stationary Environments, Applied Ergonomics, Volume 129, pp. 104598, doi: 10.1016/j.apergo.2025.104598.

Lim, Y. Ramasamy, S. Gardi, A. Kistan, T. Sabatini, R. (2018). Cognitive Human-machine Interfaces and Interactions for Unmanned Aircraft, Journal of Intelligent and Robotic Systems, Volume 91, pp. 755–774. doi: 10.1007/s10846-017-0648-9.

Liu, C. Zhang, C. Sun, L. Liu, K. Liu, H. Zhu, W. Jiang, C. (2023). Detection of Pilot's Mental Workload Using a Wireless EEG Headset in Airfield Traffic Pattern Tasks, Entropy, Volume 25, No. 7, p. 1035, doi: 10.3390/e25071035.

Liu, Y. Yu, Y. Tao, H. Ye, Z. Wang, S. Li, H. Hu, D. Zhou, Z. Zeng, L.-L. (2025). Cognitive Load Prediction from Multimodal Physiological Signals Using Multiview Learning, IEEE Journal of Biomedical and Health Informatics, Volume 29, No. 5, pp. 3282–3292, doi: 10.1109/JBHI.2023.3346205.

Mohanavelu, K. Poonguzhali, S. Janani, A. Vinutha, S. (2022). Machine Learning-based Approach for Identifying Mental Workload of Pilots, Biomedical Signal Processing and Control, Volume 75, p. 103623, doi: 10.1016/j.bspc.2022.103623.

Nadri, C. Liu, Y. Zahabi, M. Kaber, D. Ruiz, J. Middleton, M. McKendrick, R. (2024) "Analysis of Pre-flight and Monitoring Tasks Using Cognitive Performance Modeling", Human Factors in Design, Engineering, and Computing. AHFE (2024) International Conference, USA, doi: 10.54941/ahfe1005693

Naseri, N. Parastesh, F. Ghassemi, F. Jafari, S. Perc, M. Završnik, J. (2025). Synchronization Levels in EEG Connectivity During Cognitive Workloads While Driving, Nonlinear Dynamics, Volume 113, No. 7, pp. 7243–7258, doi: 10.1007/s11071-024-10729-1.

Nikolai von, J. Johannes, K. Arnd, E. Martin, B. (2022). A Subjective One-item Measure Based on NASA-TLX to Assess Cognitive Workload in Driver-vehicle Interaction, Transportation Research Part F: Traffic Psychology and Behaviour,Volume 86, pp. 210–225, doi: 10.1016/j.trf.2022.02.012.

Pei, H. Wang, Z. Cao, J. Chen, Y. Bai, Z. (2024). A Cognitive Load Assessment Method for Fighter Cockpit Human-machine Interface Based on Integrated Multi-criteria Decision Making, Applied Soft Computing, Volume 167, p. 112287, doi: 10.1016/j.asoc.2024.112287.

Taheri Gorji, H. Wilson, N. VanBree, J. Hoffmann, B. Petros, T. Tavakolian, K. (2023). Using Machine Learning Methods and EEG to Discriminate Aircraft Pilot Cognitive Workload During Flight, Scientific Reports, Volume 13, No. 1, p. 2507, doi: 10.1038/s41598-023-29647-0.

Wang, H. Pan, T. Zhao, Y. Li, Y. Li, G. Si, H. Zhang, H. (2025). Research on Pilot Workload Identification Based on Stacking Ensemble Learning, IEEE Transactions On Aerospace and Electronic Systems, Volume 61, No. 5, pp. 13887–13903, doi: 10.1109/TAES.2025.3580708.