Predicting Pilot Workloads Using Physiological Measures

Valor Carlsen, Roger Manzi, Simon Dellinger, Tanner Craig, and Donald Koban

Department of Systems Engineering, United States Military Academy, West Point, New York 10996

Corresponding author's Email: donald.koban@westpoint.edu

Author Note: The authors would like to extend our thanks to Lockheed Martin for providing the opportunity to participate in this research. The views expressed herein are those of the authors and do not reflect the position of the United States Military Academy, the Department of the Army, or the Department of Defense. This paper was previously published and presented in the Donald R. Keith Memorial Capstone Conference at USMA in May of 2024.

Abstract: In this study, we utilized machine learning algorithms to predict pilot workload based on physiological, cognitive, and eye-tracking data collected from 7 pilots performing tasks in an unclassified F-35 flight simulator. We used the Bedford Workload Rating Scale (BWRS) to validate the high workload conditions induced during the simulated flight scenarios. We then trained models and compared their performance at predicting high workloads. Our results showed model performance was higher when classifiers were trained on individual pilots instead of on a group of pilots. We found that changes in Percent Change Pupil Size (PCPS), an eye-tracking measurement, were particularly noticeable in high vs. low-workload scenarios. This metric emerged as the most significant factor in distinctly difficult situations. These findings suggest a shift towards personalized machine-learning models for enhancing human-machine interactions in aviation through biometric and PCPS monitoring. Future work should examine a more diverse set of tasks, validated by study subjects, to assess the potential benefits of incorporating artificial intelligence (AI) assistance systems into the cockpit.

Keywords: Pilot Sensing, Workload Recognition, Machine Learning, Physiological Measures.

1. Introduction

According to the Federal Aviation Administration, 80% of aviation mishaps are caused by human error, often linked to factors such as insufficient sleep, distractions, and lapses in attention (Federal Aviation Administration, 2022). These are just a few of many factors that can contribute to cognitive degradation, lead to human error, and cause mishaps. However, we believe many of these human-error-related mishaps could be prevented by using artificial intelligence (AI) to determine when pilots need automated assistance. Furthermore, we believe that monitoring and assessing a pilot's physiological and cognitive state may help with determining when pilots are experiencing high workloads. While today's 5th generation fighters boast sophisticated capabilities to monitor the aircraft state and subsystem performance, there exists a noticeable gap in the biometric monitoring of pilots. This gap underscores the pressing need for a paradigm shift in aviation technology, with experts advocating for the integration of systems that can monitor pilots' physiological state before, during, and after flight (National Commission on Military Aviation Safety, 2020). Incorporating biometric monitoring systems into the cockpit holds the promise of proactively identifying cognitive impairment or task saturation, enhancing the potential for improved human-machine teaming between pilots and automated systems.

In a broader military context, some senior Air Force leaders envision a trajectory similar to the development of nuclear weapons and precision-guided munitions, with advances in AI and autonomy emerging as the next military offset to ensure U.S. military superiority in future conflicts. Air Force General Paul Selva, vice chairman of the Joint Chiefs of Staff, emphasized that the future of the United States air power dominance hinges on our military's ability to integrate advances in AI and autonomy into military systems and processes (Deputy Defense Secretary, 2016). Likewise, the U.S. Air Force 2030 Science and Technology Strategy states that research in cognitive science, data presentation, and human-machine interfaces is vital to optimize human-machine teaming performance (U.S. Air Force, 2019). Lastly, the current Chief of Staff of the Air Force contends that collaboration with industry partners is critical to accelerating change and innovation (Brown, Jr., 2020). These strategic visions require an ability to mitigate human error in high-stakes environments through the integration of AI assistance. By focusing on the development and application of machine learning algorithms to timely detect workload-induced cognitive shifts in pilots, our work seeks to determine the optimal moments for AI intervention, and as a result, augment decision-making processes and operational efficiency. This approach not only advances the goal of reducing human error but also contributes to the broader aim of optimizing human-machine teaming performance in critical military operations.

Our capstone team partnered with Lockheed Martin's Skunk Works human systems architects to explore the feasibility of using machine learning algorithms to detect when pilots are experiencing a high workload during flight simulation training. Our study had three primary objectives: to examine how high workload and challenging tasks affect subjective ratings, assess the impact of increased cognitive workload on biometric markers (such as brain oxygenation, heart rate variability, and PCPS), and evaluate the performance and explainability of classifiers in predicting pilot workload using PCPS, cognitive, and physiological data. Although past studies have demonstrated some success using physiological measures to detect changes in pilot workload, differences in experimental design have made it unclear if machine learning performance can generalize across a diverse set of tasks and maneuvers (Charles & Nixon, 2019). Additionally, few studies have focused on cognitive measures as a measure of workload. To address this gap, Skunk Works Labs collected physiological and cognitive metrics—specifically, heart rate variability (HRV), percent change in pupil size (PCPS), and brain activity—from 7 pilots during high and low workload scenarios in an unclassified F-35 flight simulator. We then evaluated the efficacy of several machine learning algorithms trained on this data, with the primary objective of predicting the binary variable of workload intensity: high or low.

We found that classification performance was higher when the models were trained on data from individual pilots instead of aggregated data from the group of pilots. Consistent with prior research, we found that the PCPS was an important feature in predicting workload (Murata & Iwase, 1998). Our analysis suggests that future pilot sensing research may wish to include PCPS measures and that classifiers may perform better when using an individualized approach for model training.

2. Materials and Method

The analysis presented here uses data collected from pilots (N=7) who participated in a Human-in-the-loop (HITL) assessment conducted by Lockheed Martin during the summer of 2023. The following section describes the experimental design, procedure, and our analysis approach.

2.1 Participants

A total of 8 participants were recruited for the study. However, one participant was excluded from the analysis due to their awareness of the study design, introducing a potential source of bias to their results. To diversify results, Lockheed Martin recruited pilots with different flying backgrounds; two had experience with flying F-35s, three had experience on other fighter platforms such as F-16s, F-15s, and F-18s, one pilot had no fighter experience but had experience flying instruments, and one pilot who had flown multiple platforms.

2.2 Procedure

Pilots conducted flights in an unclassified development F-35 cockpit simulator lab with the scenario set in Nellis Air Force Base, Nevada. Pilots were fitted with sensors to collect biometric data including heart, brain, and eye-tracking data. Each session took ten to fifteen minutes. Pilots were first briefed on the purpose of the experiment and how to properly don and doff the Functional Near Infrared Spectroscopy (fNIRS) device to measure the brain oxygenation variables and the Electrocardiogram (ECG) device to measure heart rate variability. The fNIRS is a field-deployable noninvasive optical brain imaging technology that measures cerebral hemodynamics in response to sensory, motor, or cognitive tasks (Harrison et al., 2013, 2014). The pilots were then briefed on the scenarios while their physiological reading output on all devices was being monitored. The participant then started the experiment by conducting a calibration flight, where they flew without any direction for at least five minutes to collect baseline data. After the baseline data was collected, the simulation was stopped so the first scenario could begin.

To ensure quality data collection, the pilots were briefed to treat any abnormalities or emergencies they would experience in the simulation as they would in real life. The first scenario was designed to be a low-workload scenario and consisted of simple tasks including taking off, flying a specified route, and landing. Between the two scenarios, the pilot was given a Bedford Workload Rating Scale (BWRS) and China Lake Situational Awareness (CLSA) scale to complete (Hicks et al., 2014). The BWRS requires pilots to subjectively rate the level of workload with the amount of spare capacity they have available to perform other tasks on a 10-point scale (1 = workload insignificant, 10 = tasks abandoned). The second scenario was designed to be a high-workload scenario. This scenario consisted of tasks to induce stress on the pilot and imposed very low visibility outside the cockpit, so the pilots had to fly the approach using instruments. The Tactical Air Navigation (TACAN) failed partway into the approach for about 30 seconds to induce stress. When the TACAN came back online, the pilots completed the airway approach and landed. The experimenters gave the pilots another BWRS and CLSA to complete at the end of the second scenario.

2.3 Measures

Our study focused on mental workload (MWL). MWL is the level of arousal or effort combined with mental capacity or cognitive resources used in completing a task (Dussault et al., 2005). This is different than task load because it includes the subjective responses of a pilot related to the task load. Task load can influence MWL based on the difficulty of the task and the number of times the task needs to be completed, but it is not the only factor that is involved in measuring MWL (Kakkos et al., 2019; Young et al., 2015). If a task is complex, it may increase MWL until a person practices and becomes familiar enough to complete it (Matthews et al., 2015).

Physiological, ocular, and cognitive measurements helped us create a detailed understanding of the pilots' mental workload. Heart Rate Variability (HRV) measures include low-frequency to high-frequency ratios (LF/HR ratio), pNN50, and root mean square of successive differences between normal heartbeats (RMMSD) (Shaffer & Ginsberg, 2017). Ocular measures include pupil size, blink rate, and blink duration. PCPS has been studied in relation to MWL before, and pupil diameter change has been observed to be higher during high workload situations (Causse et al., 2010; Kramer, 1990). Another way to detect MWL is through the measurement of oxygenated (HbO_2) blood in the brain (Causse et al., 2017). Studies have found a relationship between blood oxygenation and mental workload. Overall, increasing the workload will increase the HbO_2, making HbO_2 a viable measure of MWL (van Weelden et al., 2022). Our final dataset included thousands of observations per minute for each pilot, documenting the shifts in physiological and cognitive responses between low and high workload scenarios of the simulated flights. After confirming the accuracy of our data labels, we used machine learning algorithms to predict the binary variable of high vs. low workload scenarios.

2.4 Analysis Approach

To determine if tasks assigned to pilots during the experiment were accurately labeled, we first compared the sample means of the Bedford scale ratings for the low and high workload scenarios. We hypothesized that the Bedford scale scores would be significantly higher for the high workload scenarios. Then, we compared the sample means of each measure to determine if there were differences in physiological measures between the two scenarios. We hypothesized that in the high workload scenario, heart rate variability measures would be lower (lower LF/HF ratio, pNN50, and RMSSD), there would be larger changes in PCPS, and there would be increased brain oxygenation (higher HbO_2 delta).

To evaluate the performance of classifiers in predicting pilot workload, we compared two training methods across multiple models: a generalized method that trained models on the full dataset, and an individual-centered approach that focused on personalizing predictions for each pilot. The generalized approach trains one classifier using combined data from all pilots to identify common workload indicators, while the individualized approach uses the same data to train individual classifiers for each pilot, which results in tailored models that are sensitive to each person's unique physiological reactions. Both approaches predicted a binary response variable indicating a "Low" or "High" workload.

3. Results

After analyzing both qualitative and quantitative data, we found a significant difference between the low and high workload scenarios that confirmed the accurate labeling of the two scenarios. This confirmation gave us confidence to use these labels effectively when training machine learning algorithms. Pilots agreed that the high workload scenario was more challenging but failing the TACAN did not affect their situational awareness (Table 1). This observation was supported by the difference in mean Bedford scale ratings between the scenarios. Specifically, the mean Bedford scale ratings for the high workload scenario ($M = 7.14$, $SD = 2.48$) were significantly higher ($t = 4.63$, $p = .0006$) than those for the low workload scenario ($M = 2.71$, $SD = 0.49$). In the CLSA, one pilot scored the TACAN failure as having very good situational awareness, four pilots said they had good situational awareness, and two pilots said they had adequate situational awareness. This shows that while the TACAN failure increased the difficulty of the flight task resulting in increased mental workload, it did not significantly impact the pilots' situational awareness.

We then compared the physiological responses in the two scenarios and found noticeable differences in the pilots' physiological and cognitive reactions to different levels of workload. By conducting 2-sample t-tests, we closely examined the average biometric data for high and low workload scenarios. We found significant variations in PCPS and HbO_2 changes between the scenarios for almost all pilots, with a large t-test statistic and an extremely low p-value of less than 0.001. The difference in HbO_2 might be more pronounced in a real flight scenario, as suggested by a previous study that found that greater differences in HbO_2 when measured in a real flight versus a simulator (Gateau et al., 2018). We did not observe large differences in the HRV measures, perhaps because the simulated environment failed at replicating real-world conditions.

Table 1. Bedford Workload Rating Scale (BWRS) and China Lake Situational Awareness Scale (CLSA) Ratings by Pilot. Pilots agreed that the high workload scenario was more challenging but failing the TACAN did not affect their situational awareness.

	Workload Rating		
Pilot	Scenario A	Scenario B	Situational Awareness Rating
1	3	9	very good
2	3	4	good
4	2	6	good
5	3	5	adequate
6	2	10	good
7	3	10	adequate
8	3	6	good
M(SD)	2.71 (.49)	7.14 (2.48)	

Following the significant differences observed in physiological measures between high and low workload scenarios, we explored the predictive capability of models trained on these data. To do this, we compared individualized and generalized modeling approaches to see how well they could classify these scenarios based on pilot physiological and cognitive data. As summarized in Table 2, the individualized approach, particularly using the Random Forest model, demonstrated superior performance over the generalized approach. The Random Forest model achieved high metrics in accuracy (0.91 ± 0.05), precision (0.91 ± 0.05), and recall (0.93 ± 0.04), suggesting robustness in predicting workload levels effectively.

Table 2. Performance metrics of individualized and generalized models in predicting pilot workload. Individualized approaches, particularly using the Random Forest model, demonstrated superior performance over the generalized approaches.

Approach	Model	Accuracy M (SD)	Precision M (SD)	Recall M (SD)
Individualized	Random Forest	.91 (.05)	.91 (.05)	.93 (.04)
	KNN	.90 (.06)	.89 (.05)	.92 (.04)
	SVM	.86 (.08)	.85 (.09)	.90 (.05)
	Decision Tree	.86 (.07)	.86 (.08)	.89 (.07)
	Logistic Regression	.79 (.11)	.80 (.11)	.79 (.10)
	LDA	.77 (.10)	.78 (.10)	.79 (.08)
Generalized	SVM	.76 (.09)	.77 (.10)	.80 (.08)
	Random Forest	.76 (.11)	.74 (.11)	.85 (.11)
	Decision Tree	.73 (.11)	.74 (.12)	.77 (.13)
	KNN	.72 (.08)	.73 (.09)	.75 (.07)
	Logistic Regression	.69 (.09)	.72 (.13)	.74 (.15)
	LDA	.66 (.08)	.70 (.13)	.74 (.13)
Baseline	Dummy Classifier	.51 (.07)	.58 (.06)	.57 (.08)

Lastly, plotting a feature importance plot allowed us to understand how the changes in workload were reflected in the different predictive features. Percent Change Pupil Size (PCPS) ranked highest among the predictive features of workload changes for some pilots. The data showed significant individual variations, as PCPS did not remain the most predictive feature when pilots failed to observe a large difference in difficulty between scenarios. This variation might be due to PCPS's sensitivity to light and other external influences, which could account for its fluctuating predictive power. This underscores PCPS's value in reflecting workload, while also pointing to the importance of individual variability in response to task demands.

4. Discussion

Our goal was to assess the effectiveness of machine learning algorithms in making predictions of increased mental workload in pilots. Our findings suggest that the variation in pilots' PCPS was the most significant measurement, hence, incorporating PCPS tracking could enhance future pilot sensing research. Furthermore, our results suggest that classifiers might yield improved performance by adopting a personalized approach to model training.

Our results are consistent with past studies that have shown PCPS to be highly correlated with error rates and indirectly correlated with increases in workload (Causse et al., 2010; Recarte & Nunes, 2003). However, our results have some limitations due to how the PCPS data was collected. During the data analysis phase, correlations between the pilot's verbal statements and PCPS data showed there could be biases where the pilot is looking down at their instruments, and the Random Forest is predicting PCPS as "looking down." There are also questions concerning the height of the pilot and the amount of light present in the room which could both easily affect the validity of the PCPS data recorded, and its use in a non-simulation setting (Mathôt & Ivanov, 2019). Additionally, our analysis approach does not account for temporal aspects of the data. Future studies should explore time-series-based approaches to detect changes in PCPS trends.

Qualitative feedback from the pilots revealed pilot reservations towards high levels of automation, validating Lockheed's approach to offer customizable automation levels. This suggests the need to balance pilot preferences with the efficacy and safety of automation, especially when a pilot's automation preferences might impact operational efficiency. Additionally, we benefited from collecting pilots' feedback at the end of the simulation. Our use of Bedford ratings to validate workload-inducing conditions adds methodological rigor to our study, offering a validated scenario for future research on MWL in aviation. In their exit interviews, most pilots stated that an additional warning would increase their mental workload. However, two pilots stated they wouldn't mind a warning, saying it may be beneficial if the warning gets gradually louder.

Given the observed individual differences in the PCPS and its varying predictive value, future work should emphasize the development of personalized models and predict the workload on a Bedford scale. The improved accuracy in predicting whether the pilots are in a high or low-workload scenario when using models tailored to each pilot's unique data patterns suggests a promising direction. Custom models could account for personal baseline measurements, individual physiological responses, and distinct coping mechanisms to stress. Moreover, incorporating additional context-sensitive variables that affect pupil size, such as ambient lighting conditions and external stressors, could enhance the robustness of the predictive framework. Continuous adaptation and learning from each pilot's data over time could result in a highly nuanced and dynamic model capable of real-time, accurate workload assessment.

The goal is not only to improve the detection of high versus low workload scenarios but also to enhance our understanding of pilots' workloads in real time. By doing so, we can develop intelligent assistance systems that can dynamically adjust to each pilot's current cognitive state. Such systems could offer timely interventions, optimizing task allocation and decision support, thereby ensuring better performance and safety. Continually evolving these models with more granular data will enable a more sophisticated and individualized approach to managing pilot workload in various flight conditions.

5. Conclusion

In summary, we found machine learning algorithms performed better when trained and tested on individual results, rather than generalized results. Additionally, because some pilots are hesitant about incorporating AI assistance in the cockpit, our results suggest that the ability to down-select the amount of active AI assistance will be crucial for the e-Pilot program (Higginbotham & Skaff, 2023). These results are significant for a variety of reasons. First, training models on individualized data produce an AI catered to individual pilots. By focusing on individualized data, the opportunity to refine the model increases by adding physiological data taken from personal health trackers like smartwatches. Second, the ability to down-select the AI to the desired amount of assistance is crucial for the wide variety of tasks pilots perform. Some maneuvers, such as take-off and landing, induce a lower workload than other tasks. Pilots may be more open to using AI assistance in low-workload environments where they have attention to spare. In higher workload tasks, such as a TACAN failure, decreasing the amount of assistance could help the pilots better focus on the task at hand while still receiving the most important alerts from the e-Pilot. Third, if a pilot does not trust the recommendations that the e-Pilot provides, then the program is more a hindrance than a help and can further distract the pilot from the task at hand. Lastly, future research should reconsider the prediction methodology. Specifically, predicting a binary variable such as low/high or easy/hard may not be the best way to classify the pilot's physiological data. Instead of predicting a binary variable, predicting the value of the Bedford scale, or other similar scale could yield more accurate results.

6. References

Brown, Jr., C. (2020, August). *CSAF 22 Strategic Approach—Accelerate Change or Lose*. U.S. Air Force.

Causse, M., Chua, Z., Peysakhovich, V., Del Campo, N., & Matton, N. (2017). Mental workload and neural efficiency quantified in the prefrontal cortex using fNIRS. *Scientific Reports*, *7*(1), Article 1.

Causse, M., Sénard, J.-M., Démonet, J. F., & Pastor, J. (2010). Monitoring Cognitive and Emotional Processes Through Pupil and Cardiac Response During Dynamic Versus Logical Task. *Applied Psychophysiology and Biofeedback*, *35*(2), 115–123.

Charles, R. L., & Nixon, J. (2019). Measuring mental workload using physiological measures: A systematic review. *Applied Ergonomics*, *74*, 221–232.

Deputy Defense Secretary. (2016). *Deputy Secretary: Third Offset Strategy Bolsters America's Military Deterrence*.

Dussault, C., Jouanin, J.-C., Philippe, M., & Guezennec, C.-Y. (2005). EEG and ECG changes during simulator operation reflect mental workload and vigilance. *Aviation, Space, and Environmental Medicine*, *76*(4), 344–351.

Federal Aviation Administration. (2022, March 29). *Glider Flying Handbook*. Federal Aviation Administration.

Gateau, T., Ayaz, H., & Dehais, F. (2018). In silico vs. Over the Clouds: On-the-Fly Mental State Estimation of Aircraft Pilots, Using a Functional Near Infrared Spectroscopy Based Passive-BCI. *Frontiers in Human Neuroscience*, *12*.

Harrison, J., Izzetoglu, K., Ayaz, H., Willems, B., Hah, S., Ahlstrom, U., Woo, H., Shewokis, P. A., Bunce, S. C., & Onaral, B. (2014). Cognitive Workload and Learning Assessment During the Implementation of a Next-Generation Air Traffic Control Technology Using Functional Near-Infrared Spectroscopy. *IEEE Transactions on Human-Machine Systems*, *44*(4), 429–440.

Harrison, J., Izzetoglu, K., Ayaz, H., Willems, B., Hah, S., Woo, H., Shewokis, P. A., Bunce, S. C., & Onaral, B. (2013). Human Performance Assessment Study in Aviation Using Functional Near Infrared Spectroscopy. In D. D. Schmorrow & C. M. Fidopiastis (Eds.), *Foundations of Augmented Cognition* (Vol. 8027, pp. 433–442). Springer Berlin Heidelberg.

Hicks, J. S., Durbin, D. B., Morris, A. W., & Davis, B. M. (2014). *A Summary of Crew Workload and Situational Awareness Ratings for U.S. Army Aviation Aircraft:* Defense Technical Information Center.

Higginbotham, K., & Skaff, M. (2023, December). *e-Pilot: Even Great Pilots Need a Cognitive Aid [White Paper]*. Lockheed Martin Aeronautics.

Kakkos, I., Dimitrakopoulos, G. N., Gao, L., Zhang, Y., Qi, P., Matsopoulos, G. K., Thakor, N., Bezerianos, A., & Sun, Y. (2019). Mental Workload Drives Different Reorganizations of Functional Cortical Connectivity Between 2D and 3D Simulated Flight Experiments. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *27*(9), 1704–1713.

Kramer, A. F. (1990). *Physiological Metrics of Mental Workload: A Review of Recent Progress*. Defense Technical Information Center.

Mathôt, S., & Ivanov, Y. (2019). The effect of pupil size and peripheral brightness on detection and discrimination performance. *PeerJ*, *7*, e8220.

Matthews, G., Reinerman-Jones, L. E., Barber, D. J., & Abich, J. (2015). The Psychometrics of Mental Workload: Multiple Measures Are Sensitive but Divergent. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *57*(1), 125–143.

Murata, A., & Iwase, H. (1998). Evaluation of Mental Workload by Fluctuation Analysis of Pupil Area. *Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Vol.20 Biomedical Engineering Towards the Year 2000 and Beyond (Cat. No.98CH36286)*, *6*, 3094–3097.

National Commission on Military Aviation Safety. (2020, December 1). *Report to the President and the Congress of the United States*. National Commission on Military Aviation Safety.

Recarte, M. A., & Nunes, L. M. (2003). Mental workload while driving: Effects on visual search, discrimination, and decision making. *Journal of Experimental Psychology: Applied*, *9*(2), 119–137.

Shaffer, F., & Ginsberg, J. P. (2017). An Overview of Heart Rate Variability Metrics and Norms. *Frontiers in Public Health*, *5*, 258.

U.S. Air Force. (2019). *U.S. Air Force 2030 Science and Technology Strategy*.

van Weelden, E., Alimardani, M., Wiltshire, T. J., & Louwerse, M. M. (2022). Aviation and neurophysiology: A systematic review. *Applied Ergonomics*, *105*, 103838.

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