The Value of Cognitive Workload in Machine Learning Predictive Analytics

Amy Dideriksen[1], Joseph Williams[2], Thomas Schnell[3], Gianna Avdic-McIntire[2]

[1]Collins Aerospace, Orlando, FL
[2]Collins Aerospace, Cedar Rapids, IA
[3] University of Iowa Operator Performance Lab, Iowa City, IA

amy.dideriksen@collins.com, joseph.williams@collins.com, thomas-schnell@uiowa.edu, gianna.avdic@collins.com

ABSTRACT

The Department of Defense plans to spend $1.7 billion over the next five years to stand up a new Joint Artificial Intelligence Center with goals to develop strategic plans, adopt and transition artificial intelligence, machine learning and emerging technologies into operational use (Longwell, 2018). Until roadmaps have matured, it is unclear how much of that budget will go towards training.

Several commercial industries have implemented solutions using data analytics to improve operations. Many in the military training industry are beginning to design architectures and plan research studies using learning analytics to predict performance, personalize and adapt training to optimize human performance. Large amounts of training data sets for effectively training the networks is one of the biggest challenges. Few researchers have assessed machine learning solutions that include physiological metrics to adapt learning. Our collaborative research team capitalized on the unique data collection from previous privately funded research conducted over the past two years. This research collected data on 30 pilots flying multiple flight maneuvers in a simulator and in a live aircraft with over 50 plus hours of live flight time. We collected metrics on task performance and cognitive workload.

The purpose of this study was to develop several deep neural networks to predict future states of student performance. We compared the results of our predicted performance using only task performance measures with the results of task performance measures in combination with cognitive workload metrics. The results show when cognitive workload is included in our deep neural networks; it increased the performance prediction to an extremely high level of accuracy.

ABOUT THE AUTHORS

Amy Dideriksen, PMP is a Global Training Research Manager in Mission Systems at Collins Aerospace with over thirty years of training experience and a background in Instructional Systems Design. She has an MS in Industrial Technology specializing in Training and Development/Human Resource Development and is currently pursuing her doctorate in Industrial Engineering with a focus in Human Factors from the University of Iowa. She is the Manager in Advanced Technology for research initiatives in Cognitive State Assessment and Training Effectiveness. Her interests are in integrating digitally enabled technologies with simulation-based training solutions.

Joseph Williams is a Senior Software Engineer in Collins Aerospace's Advanced Technologies Data Analytics group. He has more than 10 years of software development experience and holds bachelor and master's degrees in computer science from South Dakota School of Mines and Technology. In addition to Adaptive Learning, he is currently doing machine learning research for the Global Convective Weather product. His research interests include artificial intelligence and big data analytics.

Dr. Tom "Mach" Schnell is a Professor in Industrial and Mechanical Engineering with a specialization in Human Factors/Ergonomics at the University of Iowa. He is also the director and chief test pilot of the Operator Performance Laboratory (OPL). Tom has secondary appointments as professor in the departments of Electrical Engineering and Neurology. He has an undergraduate degree in EE and an MS and Ph.D. in Industrial Engineering with a specialization in Human Factors. He has lead over 282 human factors research projects with external funding of around $23 million and over 2,200 incident free flight test hours. He is a Commercial pilot with over 6,200 flight hours, research test pilot,

and flight instructor with helicopter, jet, and glider ratings and has been actively involved in flight-testing using fixed wing and helicopter testbeds as a pilot and principal investigator since 2004. He is also an Unmanned Aircraft Systems (UAS) pilot with Beyond Visual Line of Sight (BVLOS) flight time. At OPL, he conducted research into the training effectiveness and skill development of UAS flight crew using cognitive, mission specific, and behavioral measures to assess the training readiness of remote pilots and sensor operators.

Gianna Avdic-McIntire, MBA is a Senior Project Manager for Data Analytics in Mission Systems at Collins Aerospace. She works closely with the Advanced Technologies on research efforts. Her technical background is in statistical techniques. She is currently a Doctoral candidate at St. Ambrose University in Davenport, IA at the College of Business, with interests in Organizational Behavior and Strategy.

The Value of Cognitive Workload in Machine Learning Predictive Analytics

Amy Dideriksen[1], Joseph Williams[2], Thomas Schnell[3], Gianna Avdic-McIntire[2]

[1]Collins Aerospace, Orlando, FL
[2]Collins Aerospace, Cedar Rapids, IA
[3] University of Iowa Operator Performance Lab, Iowa City, IA

amy.dideriksen@collins.com, joseph.williams@collins.com, thomas-schnell@uiowa.edu, gianna.avdic@collins.com

INTRODUCTION

In the midst of our third boom for artificial intelligence (AI), we are seeing rapid growth in program development, strategic initiatives and continued investments from the Department of Defense and commercial industries (Richbourg, 2018). We are also seeing a prevalence in data analytics and continued growth in computer power that has allowed us to reach surprising performance levels using deep learning and machine learning in several industries such as healthcare, finance, social media, advertising, engineering, image and speech recognition, robotics and transportation (Woodard, 2018).

Machine learning has been used in many industries for operational use, but machine learning research in the training community is still in its early stages. It is believed that by enhancing training in real-time, through adaptive learning techniques, training facilities can personalize training to increase the quality and speed of learning. The industry is beginning to include human system metrics in the adaptive learning solution. With the increased interest in big data, we have seen an increase in commercial sensors to collect and report data in real-time. However, understanding what data to collect and how to appropriately use that data is still being proven through evidence-based research.

In 2017 and 2018, we validated our approach to measuring training effectiveness by assessing cognitive workload measures in combination with task performance metrics (Hoke, 2017; Dideriksen, 2018). We believe that using this same objective performance assessment metrics, we could personalize training to meet the needs of the student. Task performance metrics reflect the student's ability to perform the learning objective, and are collected by the simulation system. These are traditional performance measures that are collected and assessed by the training industry today. Cognitive workload measures are collected through electrocardiogram (ECG) waveforms and represent the student's level of engagement during the training exercise. By predicting student performance in real-time, we can modify training content to maintain student engagement, which will aid in the transfer and retention of requisite skills. We hypothesized that by including cognitive workload measures in combination with task performance metrics, our prediction accuracy would increase.

The purpose of this paper is to describe our process of collecting appropriate data, analyzing the data, and developing a supervised learning algorithm using Deep Neural Networks to predict student performance into the future. To ensure trust in where the data originated, we used the data set from the historical performance results collected from our research to train the networks.

METHOD

Our data was collected from two years of research in 2017 and 2018. The participants, testbed, measures and procedure below describe the method used to collect the data set for training the machine learning algorithms. Prior to performing flight tests, pilots conducted Perceptual-Cognitive sessions, baselining from home. The remainder of the research was conducted at the University of Iowa, Operator Performance Laboratory (OPL).

Participants

In total, 30 pilots were evaluated for this study. Each volunteer held a valid US pilot certificate, at least a current class III medical, and the mandated flight hours necessary for manually flying fixed-wing aircraft.

Testbed

The participants were seated in the rear crew position of a Vodochody L-29 jet trainer. They performed flight maneuvers with NeuroTracker operating on the upper display in the aft cockpit (Figure 1). The L-29 jet trainer was used as both an "aircraft-in-the-loop" (AIL) simulator and as a live asset. The aircraft was equipped with the Cognitive Assessment Tool Set (CATS) software that collects flight technical and physiological data from the sensors and simulation system within the testbed (OPL, 2014). Data was time-stamped, synchronized, and recorded to a database that supports data analysis.

Figure 1. Vodochody L-29 AIL, Live Asset and with NeuroTracker Touch-Screen Display

Measures

We collected both subjective (measures influenced by the observer's personal judgement) and objective (those that involve an impartial measurement, that is, without bias or prejudice) during the study. For subjective measures, pilots were asked to complete self-assessment surveys: Situational Awareness Rating Technique (SART) and a 10-point Bedford Workload scale to assess their workload during the exercise. Objective measures include perceptual-cognitive measures produced by NeuroTracker and physiological, eye tracking and flight technical measures collected by CATS.

Perceptual-Cognitive Measures. NeuroTracker is a perceptual-cognitive measurement technique based on Three Dimensional Multiple Object Tracking (3D-MOT) to stimulate a high number of brain networks that must work together during the exercise (Faubert & Sidebottom, 2012). The activity requires students to employ complex motion integration, dynamic, sustained and distributed attention processing and working memory by tracking four targets among eight spheres following a linear trajectory projected within a cube space, as illustrated in Figure 2.

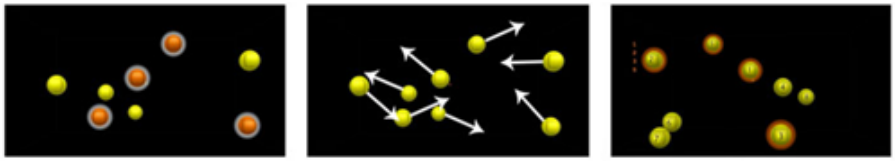

Figure 2. NeuroTracker: Target Identification, Displacement and User Response

NeuroTracker provided a secondary task used to quantify spare cognitive resources based on the displacement thresholds. Displacement thresholds represent how fast the balls are moving, where faster speeds are indicative of better performance.

Physiological Measures. Eye-tracking data was collected using a Dikablis eye tracker installed on the visor above the helmet. Eye gaze data was used to analyze the percentage of time the pilot was viewing NeuroTracker in the upper display or flight instruments in the lower display. Real-time cognitive workload estimation was generated from electrocardiogram (ECG) signals by the CATS software using a minimally intrusive system with body-worn electrodes and the NeXus-4 biofeedback device which do not interfere with performance tasks. The ECG waveform is transformed from its scalar space to an embedded phase space, then is coarse grained to provide a quantitative signature of cognitive state (Engler et al., 2013; Schnell & Engler, 2014).

Flight Technical Measures. Task performance data collected in CATS included aircraft performance metrics benchmarked against ideal or standard thresholds established for each maneuver. These measures included altitude (ft), roll (deg) and vertical speed (ft/min) errors.

Procedure

Simulator and Flight Test: Pilots received a briefing of the study objectives and an in-depth review of the flight maneuvers and safety training. Upon completion of safety training, researchers attached ECG leads. Flight maneuvers were flown in the AIL simulator configuration, followed by a live flight in the L-29 jet trainer. Each pilot performed three levels of flight maneuvers, summarized in Table 1, to assess their flight technical performance while using the perceptual-cognitive software as a secondary task to provide a measure of spare cognitive resources. Each participant also flew the three flight maneuvers without the perceptual-cognitive software present. The order of these scenarios was randomized for both the simulator and flight portions of the study.

Table 1. Flight Maneuvers

Difficulty	Right Turn	Altitude	Heading	Left Turn	Altitude	Angle of Bank
Low	360°	From starting alt, maintain ±200 ft	Reach original heading, ±10°	360°	Maintain ±200 ft and ±100 ft	30°
Medium	360°	From starting alt, maintain ±100 ft	Reach original heading, ±10°	360°	Maintain ±100 ft	45°
High	3 deg/sec to reach original heading in 120 sec	1000 ft climb at 500 ft/min	Reach original heading. ±10°, in 120 sec	3 deg/sec to reach original heading in 120 sec	-500 ft/min descent to initial altitude	Varies

DATA ANALYTICS

With the large amount of recorded pilot cognitive state and flight technical data available, simple human-observation of the data would prove to be nearly impossible and algorithmic data analysis techniques needed to be implemented. Our goal was to demonstrate that pilot behavior is relatively predictable. We also demonstrated that the addition of a reliable cognitive workload metric to the flight state vector enhances our predictions.

Data Collection. The same flight technical and cognitive workload metrics were collected during training exercises for both simulator and live flight events. Onboard data collection was facilitated by CATS, which collects data from various sensor and simulator systems (e.g., aircraft state), creates timestamps, synchronizes and records the data into a MySQL database. Sensor and simulator data devices produce data at different rates and utilizes proprietary data formats. CATS uses custom-defined providers that collect and collate the data from these variant sources. CATS allows the experimenter to effectively record and annotate data from human-in-the-loop (HITL) studies in a single, simple-to-operate user interface. Due to the influence of the dynamic real-world flight environment, data collection had real-world artifacts that needed to be removed before the machine learning algorithms could use the data. For example, some maneuvers had to be aborted and then repeated due to real-world air traffic control (ATC) interventions. Researchers familiar with the maneuvers cleaned the data and ensured that the timestamps surrounding all maneuvers were precise. For our analysis, we did not use subjective measures or eye tracking data.

Data Classification. We were faced with the challenge of flight technical data values being collected many times per second, though not always at a consistent rate. For inputs into our grading and machine learning models, data interpolation was employed to give ten evenly spaced samples per second. Because the original data sample rate was often near, and sometimes greater than 10 Hz, linear interpolation was used. Visual inspection of graphs comparing the interpolated values to the actual data confirmed that no spikes in data were being removed and that the interpolated values accurately represented the true data values.

Cognitive workload is calculated at a lower rate, but the same linear interpolation technique was used to produce data at the same rate as the other flight technical data. Again, visual inspection was used to confirm that this interpolated data accurately represented the provided source data.

Scoring

To predict a pilot's future performance, we first had to define a grading system for each flight maneuver. The low and medium maneuvers involved keeping the plane in a specific degree of roll while maintaining their starting altitude. The high maneuver involved climbing at a defined rate while turning to a defined heading in a specified amount of time, then reversing the process by descending and turning in the opposite direction. The challenge came in getting to those values as smoothly as possible.

Due to the differences in the flight maneuvers, one grading system was used for the low and medium maneuvers while a similar, but distinct, system was used for the high maneuvers.

Low and Medium Scoring. The metrics used to grade performance were roll and altitude differences from the ideal. To avoid the grade jumping dramatically, we based their score upon the last five seconds of performance. This allowed for a gradual change as the pilot's performance changed.

One of the biggest challenges to overcome was determining a scoring model. After exploring several different models and continuous conversations with the flight experts from OPL, the grading model shown in Equation 1 was adopted. It scales grades between 0 and 1, with 1 being the best a pilot could score. The ideal roll in the equations is determined by the maneuver in question, 30 and -30 for low, 45 and -45 for medium.

Equation 1:

$$\text{rollDiff} = \text{abs}\left(\text{abs}\left(\text{roll}_{\text{actual}}\right) - \text{abs}\left(\text{roll}_{\text{ideal}}\right)\right)$$

$$\text{altDiff} = \text{abs}\left(\text{altitude}_{\text{actual}} - \text{altitude}_{\text{starting}}\right)$$

$$\text{currentGrade} = \left[1000 - \left(\text{rollDiff}_{\text{90th percentile}} \times 50\right) + \text{altDiff}_{\text{90th percentile}}\right] / 1000$$

This grading system isn't without its flaws, the most obvious being that it is impossible for a pilot to get a perfect score. However, we found that it was effective at differentiating between pilot performance, changed gradually as pilot performance changed, and properly scaled to changes in altitude and roll.

The following figures are examples of how the grade changed in response to pilot performance for low and medium levels of difficulty flight maneuvers. The top line of the graph is the pilot's rolling grade. It changes smoothly in response to changes in the bottom two lines, which are "difference between the current and ideal roll" and "difference between the current and ideal altitude" respectively.

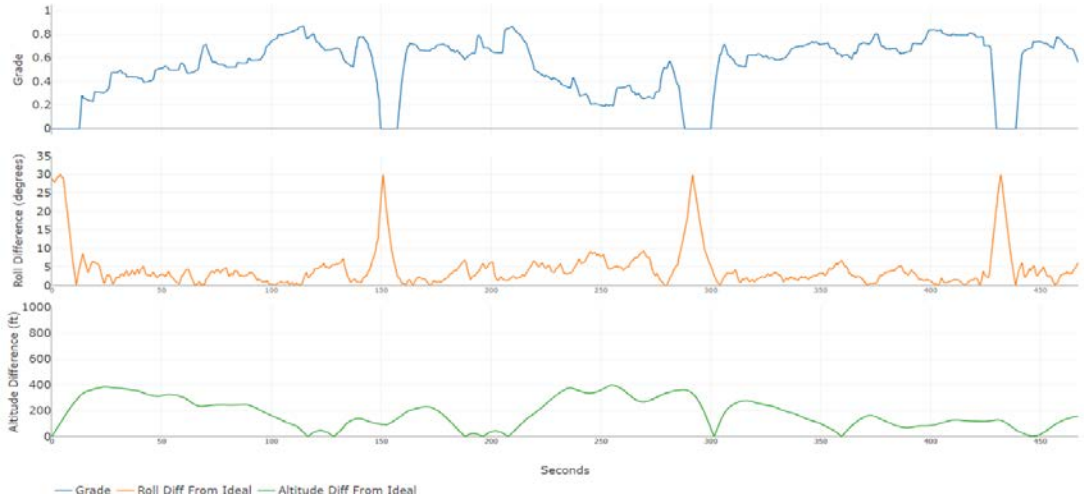

Figure 3. Low Level of Difficulty Maneuver Grading

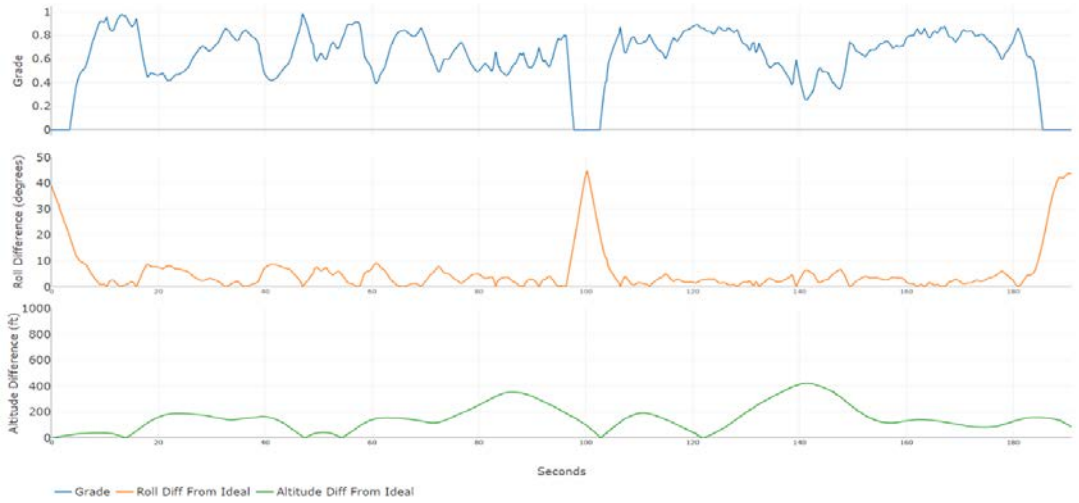

Figure 4. Medium Level of Difficulty Maneuver Grading

High Scoring. Scoring for the high maneuver followed a similar strategy as the low and medium maneuvers, however the inputs were modified to match the complexity of the maneuver, as shown in Equation 2. To grade properly, the heading and altitude were compared to an ideal state (perfectly smooth and gradual) for the maneuver. This is in contrast to the low and medium maneuvers, which compared roll and altitude to static values. Due to the increased complexity of the maneuver, the constants had to be changed to provide reasonable grades.

Equation 2:

$$\text{headingDiff} = \text{abs}(\text{currentHeading} - \text{idealHeading})$$
$$\text{altDiff} = \text{abs}(\text{currentAltitude} - \text{idealAltitude})$$
$$\text{currentGrade} = [2000 - (\text{headingDiff}_{\text{90th percentile}} \times 30) + \text{altDiff}_{\text{90th percentile}}] / 2000$$

Being consistent with the low and medium maneuvers, grades are based on the last five seconds of flight time. This allows for a gradual change in grade that incorporates both values being evaluated. Figure 5 is a graph of this grading scale for the high level of difficulty flight maneuver. The top line is the pilot's grade at a given time. The second and third lines are comparisons between the current heading and altitude and the ideal heading and altitude respectively. The ideal heading and altitude are taken from perfectly smooth versions of the maneuver.

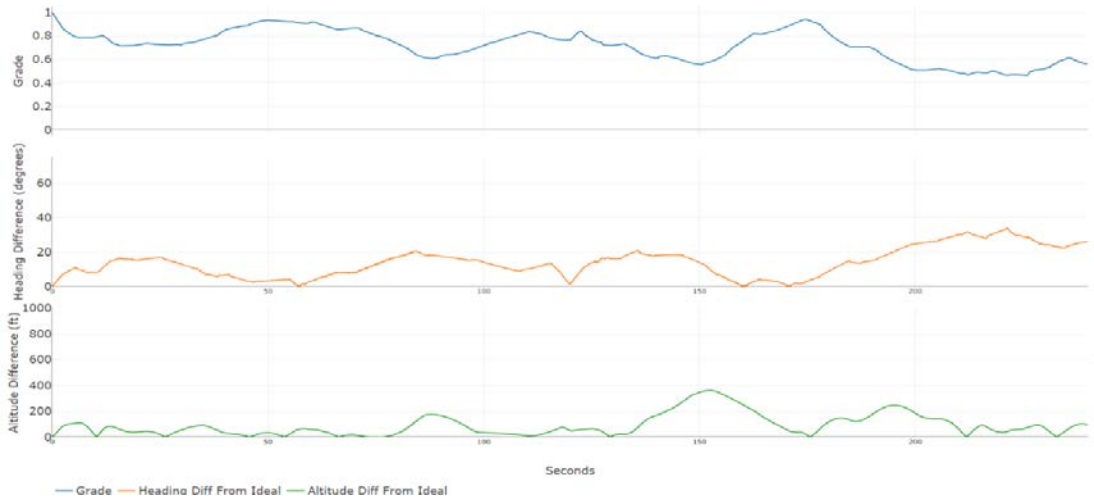

Figure 5. High Level of Difficulty Maneuver Grading

Machine Learning Steps

Determining the appropriate machine learning model was another challenge. To predict pilots' future performance based upon current time-series data, Recurrent Neural Networks (RNN) were chosen for the first attempt. We used a variation called Long Short-Term Memory (LSTM), but found that its predictive capabilities broke down a few seconds into the future. We found success using Deep Neural Networks built from 16 fully connected layers. Figure 6 illustrates the network inputs for the low flight maneuver. This network allowed us to predict 20 seconds into the future based on the past 5 seconds of performance.

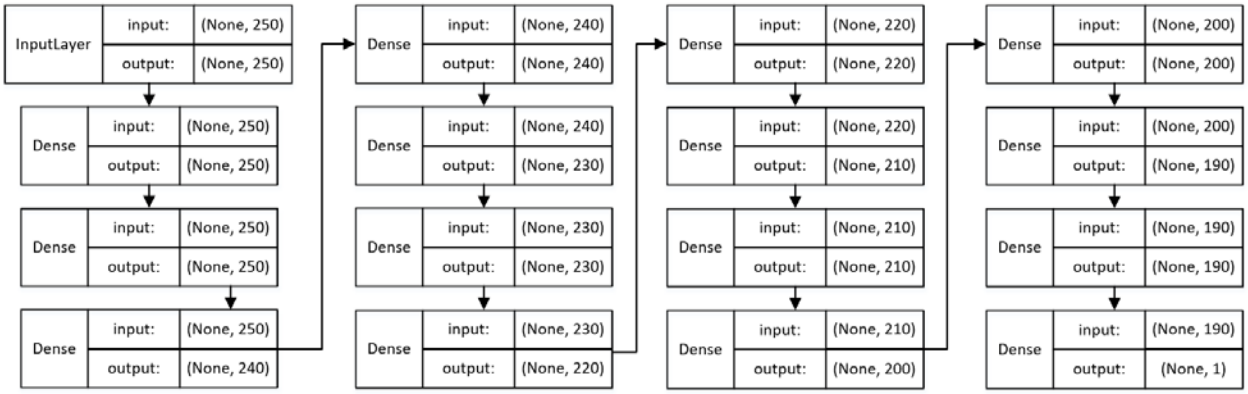

Figure 6. Deep Neural Network for Low Flight Maneuver

Training and Testing Methods

To get the most accurate representation of each network's accuracy, 10% of the data (randomly selected) was set aside before training to be used as a test dataset. The remaining 90% of the data was used for training and selecting the best network. We used a cross-validation of five within the training data to select the best network layout. This meant that each network was trained five times, each time training on 80% of the available data and testing itself against the remaining 20%. The most successful network was selected and the listed test-scores were taken from testing the trained network against the 10% of data originally set aside, which the network had never seen before.

We attempted to predict pilot performance 20 seconds into the future and provided each network with data from five seconds of flight. When attempting to predict pilot grades at time *t*, data from time *t-25* through *t-20* was provided to the network. Times are given in seconds. For each maneuver, four separate networks were trained using different sets of input variables.

score. This network was considered our baseline for the predictability of pilot scores for each maneuver. The data inputs this network used in predicting the score were the deviations from the ideal used in calculating grades.

score+CW. This network included deviations from the ideal roll and altitude, and cognitive workload.

scoreMicro. This network included deviations from the ideal roll and altitude, as well as any micro adjustments to the roll and pitch.

scoreMicro+CW. This network included deviations from the ideal roll and altitude, any micro adjustments to the roll and pitch, and cognitive workload.

RESULTS

Table 2 shows the accuracy each model was able to achieve on the test data, which is hidden from the model during training. Adding workload information consistently improved the accuracy of the machine learning models.

Table 2. Machine Learning Test Data Accuracy

Classifier	***score***	***score+CW***	***scoreMicro***	***scoreMicro+CW***
Low	87.30%	95.95%	95.61%	97.19%
Medium	83.89%	93.46%	95.06%	96.68%
High	94.94%	97.47%	98.12%	98.36%

A one-way ANOVA was performed on the results for each of the three maneuvers (classifiers). Figure 7 shows the mean for the *score* network was 88.71% accurate. This shows that pilot performance is moderately predictable using performance deviation from the ideal. The *score+CW* network results show an increase in prediction accuracy over the *score* network by 6.91% with an average of 95.62% accuracy. The *scoreMicro* network results show a significantly higher average base accuracy of 96.26%, than the *score* network, with an insignificant increase over *score+CW*. Even with this higher base accuracy, the inclusion of cognitive workload in the *scoreMicro+CW* network increase predictive performance by 1.15% with an average of 97.41%.

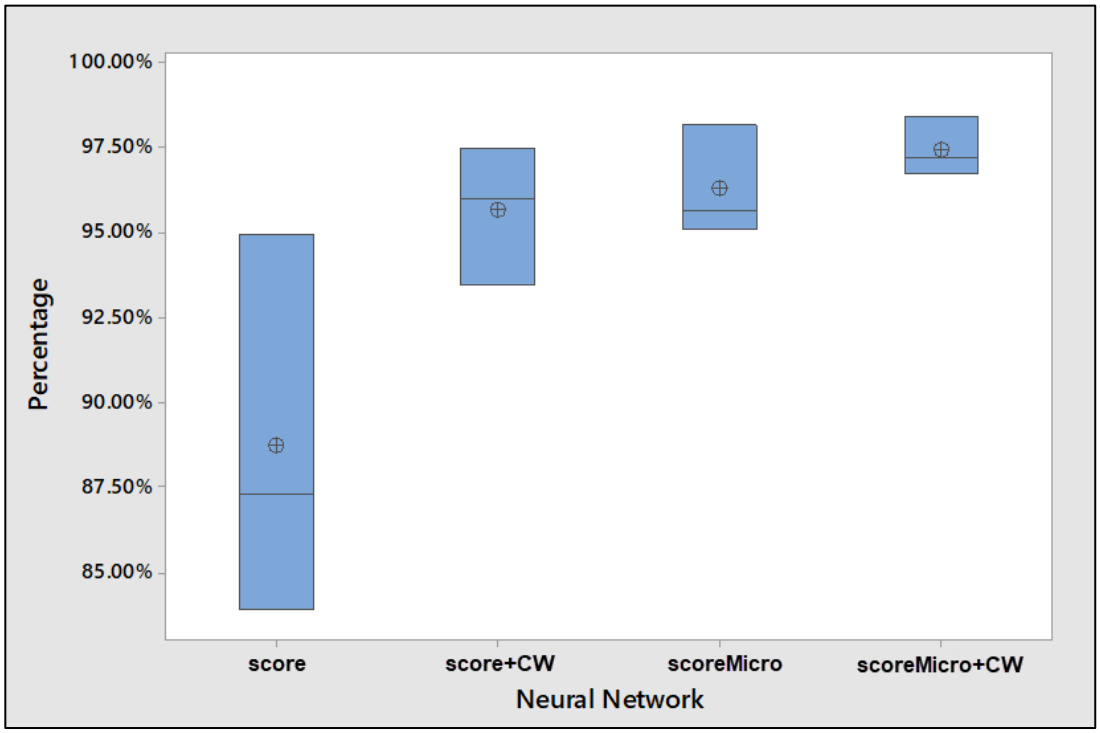

Figure 7. Boxplot of Predictive Accuracy

We conducted a statistical analysis to compare the delta (error) between the predicted grades and the actual grades for each of the three maneuvers. Using the mathematical formula in Equation 3, we computed four quantitative, continuous variables for each of the three maneuvers.

Equation 3:
Error_*score* = Predicted Score - Actual Grade
Error_*score+CW* = Predicted *score+CW* - Actual Grade
Error_*scoreMicro* = Predicted *scoreMicro* - Actual Grade
Error_*scoreMicro+CW* = Predicted *scoreMicro+CW* - Actual Grade

For each maneuver, we generated a cumulative histogram. Figure 8 represents the histogram for the high maneuver.

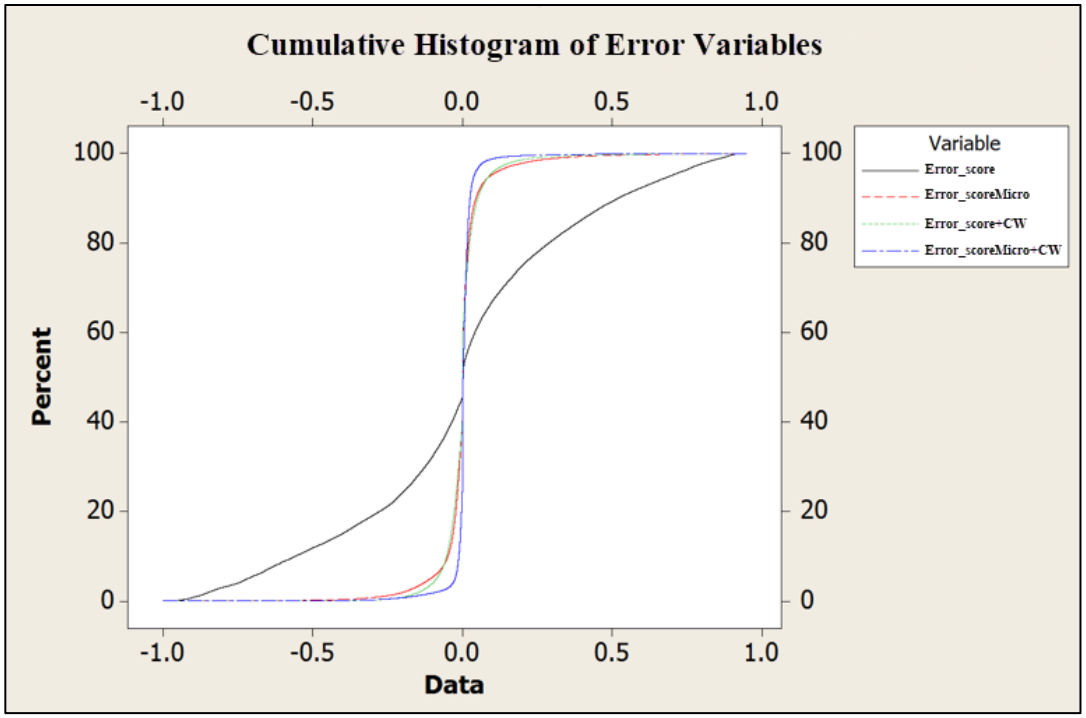

Figure 8. Histogram of High Maneuver

We conducted a Test for Equal Variances (F-test) to assess the variances between the four error variables, indicative of the four predictive networks discussed in this study. In all three maneuvers, the Test for Equal Variances confirmed the score*Micro+CW* to be consistently statistically significant.

Table 3. Test for Equal Variances (F-test)

Maneuver	N	SD	F-statistic	P-value
Low	477446	0.089466	4.12	0.0001
Medium	409014	0.087227	6.24	0.0001
High	300074	0.044048	76.04	0.0001

For each maneuver, the grades generated by predictive networks that were smoothed were consistently better predictors of actual grades. Each network takes five seconds worth of data and uses it to predict scores 20 seconds into the future. Grades are calculated from the pilots' roll and altitude deviations from the ideal roll and altitude values for the maneuver. Our baseline network uses only these values to predict future grades. We consider it a 'baseline' because it is only using the inputs into the calculated grade to predict future grades. Networks that use micro-movements are taking into account the instantaneous plane pitch and roll changes. This can be thought of in terms of changes to the plane's pitch and roll, or the size of the micro-corrections the pilot is making to maintain altitude and roll requirements. The workload networks take the baseline data and includes the pilot's workload data for those 5 seconds. Between the smoothed predictive networks, the predictive network including cognitive workload consistently performed best, and the consistently best network was *scoreMicro+CW*. Both micro-movements and cognitive workload are important predictors of immediate future pilot performance as expressed in a continuous grade.

We also ran scores and compared the results with the presence of NeuroTracker as a secondary task with NeuroTracker being absent. There was no significant difference in our overall results.

Figure 9 illustrates how the predicted score lines up with the actual performance score. In the top graph, the blue line represents the actual rolling grade from a low difficulty flight maneuver, and the orange line represents the predicted grade based on data from 20 seconds ago. In the middle graph, the green line represents the pilot's performance

deviation from the ideal for roll, and in the bottom graph, the red line represents the pilots performance deviation from the ideal for altitude.

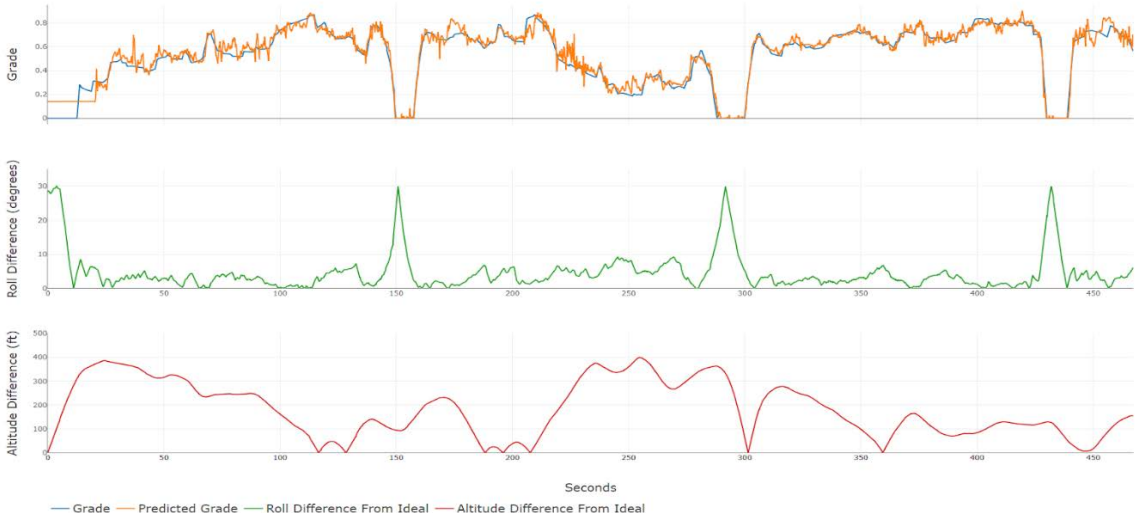

Figure 9. Low Level of Difficulty Maneuver with Predicted Grading

DISCUSSION AND CONCLUSION

The theory of using objective task performance measures to adapt learning has been around for decades, machine learning is in its third boom, and the use of physiological sensors has seen an increase in the past several years. These are not new concepts. Based on a literature review, including cognitive workload in predictive analytics is a new contribution to the training industry. Upon completion of our analysis from previous research results, we were able to predict pilot future performance with a high degree of accuracy. The overall accuracy of predicted performance increased with the inclusion of cognitive workload measures.

We did have some lessons learned from our analysis. Although these accuracy levels are great for attempting to predict pilot performance 20 seconds into the future, the predicted grade jumps around the actual grade as seen in Figure 6. By implementing a post-processing function, the peaks and valleys from the predicted grade could be smoothed to make them more useful during real-time feedback.

High level of difficulty maneuvers were easier to predict from the low and medium maneuvers, due to the nature of the maneuvers and how they are graded. It is much easier to recover from early mistakes with the low and medium maneuvers than it is with the high maneuver.

It should also be noted that there are limitations to the neural networks. The data set used was from pilots who had incentive to perform well and were trying hard to be successful. Using the neural networks to predict performance with adversarial examples, such as subjects with no flight experience or motivation for success, may provide vastly different results.

NEXT STEPS

The testing of all neural networks was retrospective and not prospective. To demonstrate that these machine learning models can actually predict future performance, live trials are needed with the addition of a post-processing function to smooth the predicted grade fluctuations. Upon completion of live trials with similar results, adaptive training solutions can be implemented. The benefit of being able to monitor and predict student performance using flight technical and physiological metrics, is that personalized training can be implemented to ensure safety and maintain engagement through an optimal level of cognitive state.

REFERENCES

Dideriksen, A., Reuter, C., Patry, T., Schnell, T., Hoke, J., & Faubert, J. (2018). Define "expert": Characterizing proficiency for physiological measures of cognitive workload. Paper presented at the Proceedings of the Interservice/Industry Training, Simulation, and Education Conference, Orlando, FL.

Engler, J., Schnell, T., & Walwanis, M. (2013). *Deterministically Nonlinear Dynamical Classification of Cognitive Workload*. Paper presented at the Interservice/Industry Training, Simulation, and Education Conference, Orlando, FL.

Hoke, J., Reuter, C., Romeas, T., Montariol, M., Schnell, T., & Faubert, J. (2017). Perceptual-cognitive & physiological assessment of training effectiveness. Paper presented at the Interservice/Industry Training, Simulation, and Education Conference, Orlando, FL.

Longwell, M. (2018). How congress wants DoD to tackle AI and machine learning in 2019. Retrieved from https://www.c4isrnet.com/newsletters/daily-brief/2018/07/24/how-congress-wants-dod-to-tackle-ai-and-machine-learning-in-2019.

Operator Performance Laboratory. (2014). *Cognitive assessment tool set (CATS) user manual*. Retrieved from Iowa City, Iowa.

Richbourg, R. (2018). *Deep learning: Measure twice, cut once*. Paper presented at the Interservice/Industry Training, Simulation, and Education Conference, Orlando, FL.

Schnell, T., & Engler, J. (2014). Entropic skill assessment of unmanned aerial systems (UAS) operators. *Journal of Unmanned Vehicle Systems*, 2(02), 53-68.

Vierling, K., Schatz, S., LaFleur, A., & Lyons, D., (2018). *Leveraging science and technology to launch innovation in learning*. Paper presented at the Interservice/Industry Training, Simulation, and Education Conference, Orlando, FL.

Woodward, T., & Enloe, M. (2018). *Deep learning applications for modeling, simulation, and training*. Paper presented at the Interservice/Industry Training, Simulation, and Education Conference, Orlando, FL