AI-augmented Human Performance Evaluation for Automated Training Decision Support

Anthony Palladino[1], Margaret Duff[1], Alexander Bock[1], Rody Arantes[1], Bernard Chartier[1], Carl Weir[1], Kendra Moore[1], Tracy Parsons[2]

[1] Boston Fusion Corp, 70 Westview St., Lexington, MA 02421, USA,
{anthony.palladino, margaret.duff, alexander.bock, rody.arantes,
bernard.chartier, carl.weir, kendra.moore}@bostonfusion.com
[2] The Royce Group, LLC, 731 Camellia Terrace Court South,
Neptune Beach, FL 32266, USA
tracyparsons@ theroycegroup-llc.com

Abstract. Human instructors must monitor and react to multiple, simultaneous sources of information when training and assessing complex behaviors and maneuvers. The difficulty of this task requires the instructor to make mental inferences and approximations, which may result in less than optimal training outcomes. We present a novel performance monitoring and evaluation system that automatically analyzes and contextualizes flight control and system data streams from high-fidelity aircraft simulators to support, validate, and augment an instructor's evaluative judgments during pilot training. We present initial results from the CAMBIO system, which leverages machine learning to assess a pilot trainee's performance in executing flight procedures. CAMBIO's machine learning approach currently achieves 80% accuracy in performance categorization.

Keywords: Flight training · Machine learning · Artificial intelligence · Human-systems Integration

1 Introduction

Flight instructors are responsible for teaching students how to operate complex machinery in high-stress and high-risk environments. These instructors must monitor the student's behavior, mindset, and response to feedback, all while ensuring the safety and stability of the aircraft [1, 2]. In military pilot training, all training must also be completed to a standardized schedule that does not allow for extra instruction for students that are struggling or an accelerated curricula for students that excel, resulting in drop-out of potential pilots and inefficiencies in training time [1]. We identified the need to develop an intelligent, adaptive system that can support, accelerate, and augment an instructor's evaluative judgments during pilot training.

In response to this need, we developed a system called Cognitive Adaptation and Management of Behavior via Information Observation (CAMBIO). CAMBIO employs methods for extracting data directly from the simulator and performing feature extraction on over 61 time-series data streams derived from steering and power controls, environmental factors, aircraft handling and stability, and other system parameters. The

features are used to train and test a series of machine learning classifiers (e.g., ridge regression and support vector machines) to categorize trainee performance and to place that trainee in relation to the population distribution of other trainees. The features are also categorized and summarized into a generalized, quantified representation of pilot behavior that we developed from literature reviews and ethnographic and exploratory data analysis investigations. Our two-pronged approach allows the utilization of advanced, but opaque, machine learning techniques for accurate categorization, while still providing understandable and interpretable results to aid in training.

2 Flight Training

Aviation instructors must keep track of dozens of variables as they evaluate their students. Accurate monitoring of this high-dimensional complex data space is not feasible by human instructors, and therefore their assessments are often subjective, based on their overall feeling of the student's performance. Computers can record these high-dimensional data streams, and CAMBIO can support instructors by providing objective measurements of pilot performance.

Emergency Procedure (Autorotation). To test our proof-of-concept system, we are currently focusing on a specific helicopter emergency procedure: autorotation. Pilots must perform the autorotation procedure when either the main rotor or tail rotor experiences a loss-of-drive and the aircraft begins to fall or spin out of control. The autorotation procedure enables the pilot to remain in control of the aircraft, navigate to a safe landing zone, and touch down safely. The procedure is divided into four main steps: entry, glide, flare, and landing [3].

Flight Simulators (Data Sources). Simulators allow both trainees and experienced pilots to practice life-saving procedures in otherwise dangerous conditions, e.g., engine failure. Additionally, flight simulators are often fully instrumented providing continuous data streams from flight controls, flight instruments, buttons, switches, audio, and the external environment.

Data Collection and Labeling. In this proof-of-concept study, we collected data from participants performing an autorotation emergency procedure in a UH-1 "Huey" helicopter simulator, built by Merlin Simulation [4]. While the participants performed the maneuvers, a trained instructor provided categorical ratings (good, fair, poor) on each of the four major steps of the autorotation (entry, glide, flare, and touchdown). In all, we collected 84 iterations of autorotation, from four pilots. Each iteration consists of 20 continuous variables (cyclic and collective positions, rate of climb, etc.), and 41 binary variables (switch positions, button presses, etc.), all collected at 60 Hz.

3 Performance Evaluation

CAMBIO takes a dual approach to classifying a trainee's performance. First, we implemented a rule-based assessment (defining rules from training manuals), described in Section 3.1. Second, we developed a Machine-Learning-based classification using our labeled simulation data, described in Section 3.2.

3.1 Rule-based Classification

Metrics for evaluating performance must be valid and have a scale [5]. The validity of our rule-based metrics is assured in that they are derived directly from the training manuals specific to the aircraft under consideration. The scale depends on the metric in question. This first stage of CAMBIO's assessment determines whether the pilot falls within the performance bounds as required by the training manual but provides limited additional quantitative feedback.

Table 1. Rules derived from training manuals for the autorotation emergency procedure's "entry" step.

Simulation Variable	Description	Lower Bound	Upper Bound	Central Value
Airspeed	Knots integrated air speed (KIAS)	80 kph	85 kph	-
Altitude	Altitude at point of entry	500 ft	None	-
Collective	Collective all the way down	-	-	0
Rotor speed	Maintain targeted speed	101%	105%	-
Pedal position	Pedals centered	-	-	50%
Rate of climb	Must not be gaining altitude	None	0 ft/sec	-
Pitch	Keep aircraft steady	-	-	0%
Roll	Keep aircraft steady	-	-	0%
Side slip	Keep aircraft steady	-	-	0%

3.2 Machine Learning Classification

Assuming the pilot performed the procedure within the bounds specified by the training manual (Section 3.1, above), we now turn to a more subtle quantification of their performance, that could be used by instructors to provide objective scoring, constructive feedback, and trend analysis by comparing with historical performance. CAMBIO contains three modules for this fine-grained assessment: (*i*) feature extraction, (*ii*) classification, and (*iii*) explainability, each described below.

Feature Extraction. The first step in developing a machine learning classifier is engineering the features that the model will use to distinguish between classes. CAMBIO uses TSFresh, a Python library for automatically calculating a large number of time series characteristics [6]. Using TSFresh, CAMBIO automatically extracts 7,200 total features per autorotation iteration (90 features / continuous variable, with 20 continuous variables / step, and 4 steps/iteration. Table 2 contains descriptions of selected TSFresh features. We use these time series characteristics as the features for our ML models.

Classification. Due to the large number of features and potential correlations between them, it is impossible to know, *a priori*, which machine learning model will achieve the best results. Additionally, it could be that different flight procedures, or steps within the procedures, will be better classified by different machine learning models. Therefore, we tested a series of models and obtained the results shown in Table 3. For autorotation's "entry" step, the Linear Support Vector Classifier achieved 80.5% accuracy in classifying a pilot's performance into the good/fair/poor categories. Figure 1 shows the confusion matrix for this trained model, showing the ability to easily distinguish between all three classes.

Table 2. Example subset of TSFresh features with statistically explainable or prominent features, descriptions (algorithmic where applicable), and examples (directly referenced in this study)

Name	Description	Example Feature
autocorrelation	Given a lag parameter, calculates autocorrelation of the time series	"Feature 382" (Fig. 1); applied to collective position data stream
cwt_coefficients	Given widths and a coefficient parameters, calculates a continuous wavelet transform of the time series	"Feature 1017" (Fig. 1); applied to rotor speed data stream
ratio_beyond_r_sigma	Given a parameter, *r*, calculates proportion of time series values greater that *r* standard deviations from the mean time series value	"Feature 858" (Fig. 1); applied to altitude data stream

Table 3. Model accuracies for classifying pilot autorotation performance

Model	Accuracy
Gaussian Process Classifier	0.683
L1 Logistic	0.683
Ridge Regression	0.756
L2 Logistic (Multinomial)	0.780
L2 Logistic (One vs. Rest)	0.780
Plain Stochastic Gradient Descent	0.780
Weighted Stochastic Gradient Descent	0.780
Linear Support Vector Classifier	0.805

ML Explainability Analysis. While the feature extraction and machine learning classification pipeline previously described offers the ability to categorize novel sets of flight simulation data, it is limited in usefulness to instructors in that it can only provide a holistic, even if confident, performance label for a given dataset. The aim of the CAMBIO system is not only to categorize performance but also to aid instructors in understanding the particular reasons performance judgments are made, allowing for more informative and fine-tuned guidance.

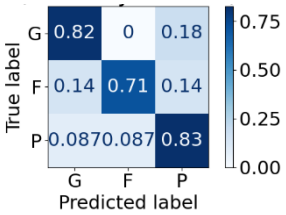

Fig 1. Confusion matrix for the Linear Support Vector Classifier with 80.5% overall accuracy

To enhance interpretability of the machine learning classifiers used, we leveraged the Shapley Additive Explanations (SHAP) tool, a Python library equipped to provide explanations for a machine learning model's decision-making. SHAP employs game theoretic techniques primarily to analyze the influence of individual features on predictions or classifications made by the analyzed machine learning model [7]. SHAP can therefore assess a model's reasoning for a given prediction based on individual features.

We used SHAP to analyze each of our model's predicted performance classes of data streams from the autorotation steps of a subset of our simulation datasets. We generated SHAP decision plots to determine the influence of each feature on the relative probabilities of the model classifying given simulation datasets as each performance label.

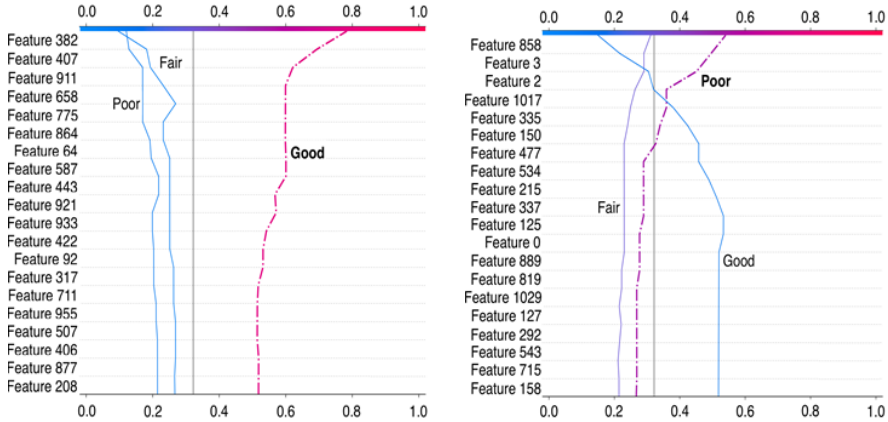

Fig. 2. Feature contribution to SVC classifications for autorotation entry steps of iteration 42 (*left*) and iteration 61 (*right*). The horizontal axis represents SHAP values converted to probabilities; the vertical axis shows the cumulative influence of the 20 most influential features on the classification result. The dashed line indicates the class to which the classified simulation iteration actually belongs.

Figure 2 shows feature contribution analysis for the performance classification of autorotation entry steps of four simulation iterations. Iterations 42 and 61 are classified into their true respective performance classes (good and poor) with moderately high confidence, indicated by SHAP probabilities of over 0.33. (The average probability of classifying a candidate into one of three classes is 0.33, indicated by the vertical grey line visible in the plots; this is treated as a baseline.) The most influential feature in the (correct) classification of autorotation entry in iteration 42 describes the data stream tracking collective position ("Feature 382"), while that of the classification of iteration 61 is a description of the data stream tracking aircraft altitude ("Feature 858"). Considering these examples, some of the most influential features statistically describe data streams of particular importance during performance of the autorotation's entry step—namely, collective position, helicopter altitude, and speed of the helicopter's rotor.

4 Use in Training

The CAMBIO system is not designed to replace instructors or remove them from the decision-making process. Instead, we have designed CAMBIO to support instructors in making faster, more confident decisions by augmenting their judgments with quantitative, tangible data and analytics. Our system provides continuous data that provides greater detail than current categorical rating paradigms, real-time performance analysis that can be used by the instructor provide feedback when necessary, and tracking over time to measure the student's progress. In addition to supporting the instructor in real-time, we also expect that the exposure of longitudinal high-resolution, high-dimensional information from CAMBIO may start to train the trainers and teach them to be more perceptive to nuances in behavior and less subjective in their evaluations.

User Interface and Data Visualization. Data and features extracted from flight simulators by the CAMBIO system needs to be intuitively and clearly communicated to

either an instructor or student. We developed two user interfaces to visualize and contextualize different relevant flight performance features. The instructor UI dashboard provides insight into the student's performance over all extracted features and over time, as well as placing the student in relation to the performance of all collected data. This provides the instructor an easy way to understand the student's current ability compared to both their own prior performance and all previous students' data. The student UI is more limited as to not overwhelm the student. This dashboard presents only a couple of features at a time by identifying an aspect of training that they are doing well on and an aspect that they should focus on improving. Students can use this dashboard to keep motivated by understanding where they are succeeding and improving their self-assessment and self-correction by understanding where they are struggling.

5 Conclusions

The CAMBIO system matched the instructor's labels with over 80% accuracy, even with our small sample size. Our results demonstrate that automated feature extraction and classification of performance can align with a trained instructor rating, while also providing a data-driven record of why and how those ratings were determined. The CAMBIO system will allow instructors to identify strengths and weaknesses of their trainees quickly and accurately, as well as to quantitatively track their performance in a multi-dimensional space over time. We will confirm these methods with a larger data collection from participants of varying experience and skill in Fall 2020. Additional future work includes automatically identifying patterns of learning over time and demonstrating the transferability of the CAMBIO system to other aircraft types.

Acknowledgments. The research and development described in this paper was sponsored by the Department of the Navy, Office of Naval Research, under contract N00014-16-C-3023. This paper is approved for public release, distribution unlimited. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Office of Naval Research.

References

1. Stein, E.S.: The Measurement of Pilot Performance: A Master-Journeyman Approach (1984)
2. Kalavsky, P., Rozenberg, R., Socha, L., Socha, V., Gazda, J., Kimlickova, M.: Methodology of Pilot Performance Measurements, Magazine of Aviation Development 5(2), 25--30 (2017)
3. U.S. Federal Aviation Administration (FAA): Helicopter Flying Handbook (FAA-H-8083-21B) - Chapter 11, Helicopter Emergencies and Hazards (2019)
4. Company website: https://merlinsimulation.com/
5. Rantanen, E.M., Talleur, D., Taylor, H., Bradshaw, G., Emanuel, T., Lendrum, L., Hulin, C.: Derivation of Pilot Performance Measures from Flight Data Recorder Information (2001)
6. TSFresh Python package documentation: https://tsfresh.readthedocs.io/
7. Lundberg, S.M., Lee, S.: A Unified Approach to Interpreting Model Predictions. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 30, pp. 4765--4774. Curran Associates, Inc. (2017)