Real-Time Situation Awareness Assessment for Pilots via Machine Learning: Constructing an Automated Classification System

Nicholas Crothers, Yash Sinha, Eric C. Larson
Southern Methodist University
Dallas, TX
{ncrothers, ysinha, eclarson} @smu.edu

Sandro Scielzo
CAE USA
Arlington, TX
sandro.scielzo@caemilusa.com

ABSTRACT

Assessing Situation Awareness (SA) for pilots in real time is critical to accelerate training of complex skills and maintain mission readiness. Traditional measures of SA are burdensome and subjective, thus cannot contribute to real time insights and adaptation in a simulation-based training environment. We have previously demonstrated that biometrics-based Machine Learning (ML) classifiers can be used to partly operationalize the construct of SA (Scielzo, Wilson, and Larson, 2020). That is, real time eye scan patterns and cognitive workload were shown to relate to level 1 SA, perception, level 2 SA, comprehension, and, to a lesser extent, level 3 SA, or projection. The research presented here represents a continuation of this study, and a first step in demonstrating the diagnostic value of real time and objective measure of SA. Specifically, we built on our existing performance data that assessed pilots' SA performing a high-altitude intercept mission, using the SA Global Assessment Technique (SAGAT) measure, by employing a mix of biometric data streams gathered from wrist-worn and eye tracker devices to regress automatically to SA scores. These biometric data streams build from our previous work in predicting scan pattern quality and cognitive load. A number of transfer learned neural network models are investigated for combining these data streams, including an investigation into the best shared network representations. We compare the performance of random forests and neural networks in both a regression and classification task across 40 subjects, achieving 77% accuracy for Level 1 and 70% for Level 2. We compare the feature importance of the trained models between Level 1 and Level 2 to evaluate whether cognitive load or gaze pattern accuracy has more importance on situation awareness. Results are encouraging for constructing an automated SA classification system. Findings are discussed in terms of expected benefits for accelerating trainees' skill acquisition and promoting warfighter readiness.

ABOUT THE AUTHORS

Nicholas Crothers is a graduate student pursuing his M.S. in Computer Science in the Bobby B. Lyle School of Engineering, Southern Methodist University. He received his B.A. in Computer Science and B.A. in Music in 2021 at Southern Methodist University. His research interests are in machine learning for education and healthcare applications.

Yash Sinha is an undergraduate student pursuing his B.S. in Computer Science in the Bobby B. Lyle School of Engineering, Southern Methodist University.

Eric C. Larson is an Associate Professor in the department of Computer Science in the Bobby B. Lyle School of Engineering, Southern Methodist University. His main research interests are in machine learning, sensing, and signal & image processing for education and security applications. He received his Ph.D. in 2013 from the University of Washington. He received his B.S. and M.S. in Electrical Engineering in 2006 and 2008, respectively, at Oklahoma State University.

Sandro Scielzo is a multidisciplinary scientist at CAE in the areas of human systems engineering, human factors, and instructional system design. Dr. Scielzo received his PhD in Applied Experimental Human Factors and M.S. in Modeling & Simulation from the University of Central Florida in 2008 and 2005 respectively. His research has concentrated on the validation and implementation of next generation training solutions for military and commercial applications to accelerate learning, guarantee proficiency, transfer knowledge, and maintain skills in the operational environment.

Real-Time Situation Awareness Assessment for Pilots via Machine Learning: Constructing an Automated Classification System

Nicholas Crothers, Yash Sinha, Eric C. Larson
Southern Methodist University
Dallas, TX
{ncrothers, ysinha, eclarson} @smu.edu

Sandro Scielzo
CAE
Arlington, TX
sandro.scielzo@caemilusa.com

INTRODUCTION

Assessing a pilot's psychological and cognitive state as a means to measure flight performance has been a major goal of the military and academia, especially in the context of developing enhanced training systems that improve readiness. Situation awareness (SA) is one such method to assess performance of pilots during training. Traditionally, measuring SA objectively is a tedious process that must be designed by trained scientists and specialists, and analysis of the results can only be done post-hoc due to the nature of the data and collection process (Endsley, 2017). The process of measurement typically requires a pilot to disrupt their training to answer questions related to the flight environment. Moreover, measuring SA in this manner requires personnel to be directly involved with the monitoring and assessment of a pilot throughout training, which is both expensive and time consuming. Automating the measurement of SA would accelerate the achievement of readiness in pilots without the need to involve more personnel in the training of pilots.

In our previous works in assessing neurological state (Wilson, Nair, Scielzo, and Larson, 2021) and flight correctness in pilots (Wilson, Scielzo, Nair, and Larson, 2020), we hypothesized that the cognitive and gaze measures could assist in the development of an automated SA measurement. Our results showed that subjective measures of cognitive load and instructor assigned gaze pattern classification together provide meaningful information towards assessing SA (Scielzo, Wilson, and Larson, 2020)—however the assessment was not completely automated. The contribution in this paper is a method to combine biometric data and eye tracking data to evaluate SA in both a regression and classification task. That is, we use machine learning to automate the classification of SA for pilots in an immersive, simulated flight scenario. We show that various levels of SA can be classified at upwards of 70% accuracy across 40 participants. All human subjects' experiments were approved by the SMU institutional review board.

BACKGROUND AND RELATED WORK

Our approach to the evaluation of situation awareness has two primary components: visual scan pattern accuracy and cognitive load of the pilot. First, we discuss how we define and assess situation awareness. We then discuss the components used to evaluate situation awareness using machine learning models.

Situation Awareness

There are many techniques that are used to assess situation awareness of pilots, each using different concepts of what situation awareness is and methodologies to assess it. The definition of situation awareness most cited and most used today is Endsley's three-level model (Endsley, 1995), which is the technique we used for data collection. This model describes situation awareness as a product of three assessment levels in a hierarchy:

- Level 1: *Perception of the Elements in the Environment*. Achieving this level of situation awareness requires the pilot to perceive the status of their aircraft and the environment around it
- Level 2: *Comprehension of the Current Situation*. Achieving this level of situation awareness requires the pilot to interpret the information gathered in Level 1 to extrapolate meaning from the situation
- Level 3: *Projection of Future States*. Achieving this level of situation awareness requires the pilot to combine elements of Level 1 and Level 2 to project information regarding the future state of their environment

Each level depends on the previous, so one would expect that, for instance, achieving Level 2 situation awareness would not be possible without first achieving Level 1 situation awareness. Further, each subsequent level requires more abstract thinking and application of visual information. For example, in Level 1, most of the information gathered is purely through observations of information present within the aircraft—such as aircraft speed, direction, and elevation—and without—such as mountains or other aircraft. This information is obtainable almost entirely using visual observation; therefore, assessing Level 1 situation awareness may be able to be achieved primarily through tracking where the pilot is looking. Achieving Level 2, on the other hand, could, for example, require a pilot to interpret the meaning of several enemy aircraft within close-range proximity to each other and their formation. Obtaining this information requires further analysis of observed visual information and would thus require assessing both gaze information as well as cognitive load information of the pilot.

The concept of different feature importance for each situation awareness level was hypothesized by Scielzo et al. (2021). In this paper, we explored the correlation between situation awareness performance, gaze pattern performance, and cognitive load during a maneuver. Based on the data gathered, we hypothesized that a pilot's performance in Level 1 would depend more on gaze pattern performance than cognitive load, performance in both Level 2 and Level 3 would depend more on cognitive load than gaze pattern performance. In our experiments, we captured these various levels of SA identically to that described in (Scielzo, Wilson, and Larson, 2020), using observation and review by expert flight instructors.

Visual Scan Pattern Accuracy

Visual scan pattern data is crucial in determining a pilot's situation awareness without directly questioning them, particularly for Level 1. Our previous work (Wilson et al., 2020) uses eye tracking data gathered from a head-mounted display to classify gaze pattern correctness of pilots during a maneuver. The data collected using this method produces a set of x, y coordinates for each eye, sampled at a frequency of up to 120 Hz. This data can be transformed into a 3-channel heatmap over a window of time, which is then passed through a VGG-based (Simonyan and Zisserman, 2015) convolutional neural network to produce a score for that gaze pattern (Wilson et al., 2020).

Cognitive Load

Mental workload is another crucial component in automatically assessing situation awareness of pilots. Our previous work (Wilson et al. 2021) uses time series biometric data gathered from a wrist-worn device to predict cognitive load based on the NASA-TLX (Hart 2006; Hart and Staveland, 1988) and Bedford (Roscoe and Ellis, 1990) systems. This method collects multiple biometric modalities as input data: photoplethysmography, electrodermal activity, wrist acceleration, and peripheral skin temperature, along with several engineered features created from these raw signals. This data is used to train a model named biometric, multi-modal, multi-task X-vector (BM^3TX) architecture (Wilson et al., 2021), which is based on variational auto-encoders (Kingma and Welling, 2013) and an X-vectors (Snyder, Garcia-Romero, Povey, and Khudanpur, 2017; Snyder, Garcia-Romero, Sell, Povey, and Khudanpur, 2018) inspired architecture.

Given the research discussed here we can formulate the following research questions: *Can we accurately and reliably use eye tracking data in conjunction with physiological/biometric data to assess situation awareness? If so, how important is each data input modality in generating this assessment as it pertains to our existing hypotheses?*

METHODOLOGY

This section describes the various data employed and models used in our work. We discuss the collected human subjects' data, data preprocessing, and model architecture, each in turn.

Human Subjects Experiment

The participant data we use for training is the same as described in our previous works (Scielzo et al., 2020; Wilson et al., 2020; Wilson et al., 2021). Forty subjects participated in the study: twenty pilots, nine operators, and eleven novices. Each subject was grouped by the amount of flight experience they had across both military and civilian aircraft; pilots had the most experience, operators had experience in roles other than a pilot—such as a combat systems officer—and novices had no flight experience.

Each subject performed two flight maneuvers for which their SA was assessed using traditional methods. This traditional assessment was used to inform the design and evaluate out machine learning models. For each event, there were two "blackout" points where the simulation was paused, the subject was asked a set of queries that pertained to knowledge of their situation, and their responses were recorded. The query responses were scored by two Instructor Pilots who determined the ground truth by reviewing the mission recordings for each subject. Both instructor pilots were former US Navy TOPGUN graduates and instructors, with more than 3,000 instructor hours between them. Each subject was asked the same set of 27 SA Global Assessment Technique (SAGAT) queries which were based on Endsley's (1988 and 2017) methodology. The two Instructors followed a two-step process for scoring: independent scoring, followed by a joint session to resolve disagreements. The outcome of this process yielded 98.25% agreement. The 27 SAGAT queries for each level and some example queries are shown in Table 1.

Table 1. Breakdown of 27 SA queries by level, with examples of each SA level

SA	**SA Measure Level**	**N**	**SA Queries Examples**
Level 1	*Perception of elements in the environment*	10	What is your altitude? What is your throttle setting?
Level 2	*Comprehension of the situation*	12	What is your vertical separation? How far off are you from desired pitch attitude?
Level 3	*Projection of future states*	5	When will you reach desired altitude? When will you reach desired offset heading?

Simulation Device

Participants used the CAE Blue Boxer^{TM} Extended Reality (BBXR) deployable training system. The BBXR is a low-footprint mixed reality training device that combines physical cockpit attributes (accurate front main panel and stick/throttle with high-precision hand tracking) with a Virtual Reality (VR) environment (see Figure 1, left). The environment was displayed via the HTC VIVE Pro Eye VR headset with integrated Tobii eye tracker, which can be worn with eyeglasses. The BBXR was selected because it maintains a high refresh rate of 90Hz, found to negate VR induced sickness (of the 40 participants, no sickness was reported), and because we have full control over the visual environment for custom manipulations of the sim. We also used a number of sensing modalities including photoplethysmography (PPG), electrodermal activity (EDA), peripheral skin temperature, and wrist acceleration in conjunction with gaze metrics including pupillary response and eye blinks. In order to keep the form factor for the biometric measurements less intrusive an all-in-one, wrist-worn device, the Empatica E4 wrist band, was selected as a data collection instrument (Khalili-Mahani, Assadi, Li, Mirgholami, Rivard, Benali, ... and De Schutter, 2020).

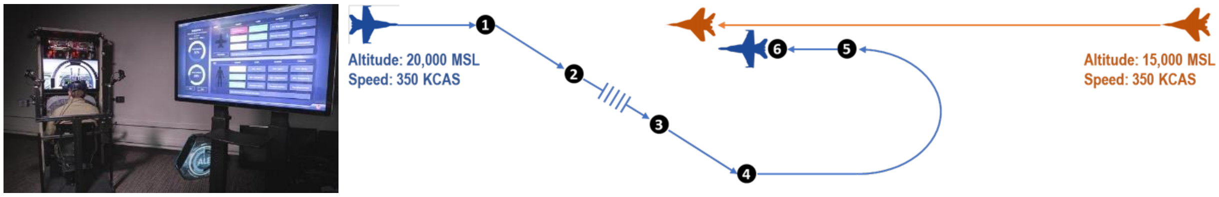

Figure 1. CAE F/A-18 Blue Boxer^{TM} Mixed Reality Device (left). Intercept profile used in experiment (right)

Air Intercept Mission

This scenario was designed to offer some challenge to experienced pilots, but to also be feasible and engaging for novices. Figure 1 above (rightmost graphic) illustrates the intercept profile, consisting of a F/A-18 blue air (flown by the participants) versus a MiG 29 red air (flown by a confederate). The tasks for the intercept mission are: (1) acquire a radar lock on bogey, (2) perform 30° offset left or right, (3) descend 10° nose low to the bogey's altitude, (4) level-off and accelerate above 400 knots, (5) perform an intercept/escort profile, and (6) perform visual identification (VID) and verification of the aircraft markings. More detailed information about the study can be found in our previous work (Scielzo et al., 2020).

Data Preprocessing

Although we attempted to collect 80 total situation awareness scores, due to equipment malfunctions (from our wrist worn biometric sensor), there are only 72 scores available. For each of the 72 scores available, there are two blackout points. The simulation is paused for several minutes each time, so only biometric and gaze data collected during a pilot's active simulation should be considered as input data. However, the time between each blackout point is separated by several minutes of unusable data, so the time series data should not be continuous throughout that period. As a result, there are theoretically two input data instances for each target score—the time before the first blackout point and the time between when the first blackout ends and the second blackout begins. For some of the maneuvers, some collected data was lost near the end of the maneuver. As a result, there are 72 input data points for the first blackout point and 65 for the second blackout point, resulting in 72+65=137 SA labeled time segments. In our analyses we use these 137 time segments for training our machine learning models. Each time segment has a series of both biometric data and eye tracking data over the associated time range. However, these data are too few to train completely new machine learning models on the raw data segments. Therefore, each segment is fed through a pretrained network from our previous works (as described below) which acts as a feature extraction model. Because our models in previous works were trained on classification tasks with many more labels, these models are more effective at reducing the time segments dimensionality. The output vectors of these feature extraction models are then used as the training data for new models of SA. This usage of previously trained models to extract features for new models is often referred to as pre-training or transfer learning (Baxter, Caruana, Mitchell, Pratt, Silver, and Thurn, 1995; Pan and Yang 2010). We give a brief description of each pretrained model from our previous works as follows:

- *Cognitive Load*: To process the time series data from the Empatica E4, we employ the *BM^{3}TX* model (Wilson *et al.*, 2021) trained for cognitive load classification as a feature extractor. This model has a multi-task output used for various cognitive load classifications. Since we are only interested in the feature extraction and not the exact measure of cognitive load, we use the vectors from penultimate neural network layer for each task and concatenate them together to produce a 112-dimensional vector. In theory, this vector contains reduced dimensionality information that is related to cognitive load.
- *Gaze Classification*: To process the gaze information from each time segment, we employ the scan pattern classifier from our previous work (Wilson *et al.*, 2020). In this model a VGG-based convolutional neural network (Simonyan and Zisserman, 2015) trained on gaze pattern data is used as a feature extractor. This model was originally used to assess if a pilot was performing the correct scan pattern in a given flight maneuver. This model has a multi-task output, so the output of the penultimate layer of each gaze classification task is concatenated together to produce a 96-dimensional vector. In theory, this vector contains reduced dimensionality information related to how a pilot scan's his or her environment.

Combined, the cognitive load features and gaze features form a set of 208-dimensional vectors (96 gaze features + 112 cognitive load features), each corresponding to a time segment with a SA score. This set of input data and targets is the dataset used for training the models to predict SA. The data can be used to train a continuous regression of SA or discrete classification of SA into various levels. In our analyses, we use a binary value for SA. That is, we assume that the primary usage of SA classification is to understand a yes/no question: *"Does the pilot have an adequate SA score, at level 1 and at level 2, for the given maneuvers?"* Because SA is a critical cognitive construct that is closely related to mission performance in complex and dynamic domains, this binary assessment is helpful for providing instructors with the ability to divide pilots into above average and below average groups. Such a grouping is essential for evaluating pilots in a pass/fail learning assessment.

For regression, the raw numeric values may be used for training a machine learning model. However, to train a classifier using the situation awareness scores, the targets must be quantized into two discrete classes: "pass" and "fail." The choice of threshold for quantization is one that splits the data into two "natural" groups according to the median score. For Level 1, the chosen split threshold was a score of 65. For Level 2, the chosen split threshold was a score of 77. Each chosen split separated the targets into a roughly equal class balance of approximately 45% one class and 55% the remaining class. This 45/55 separation was the most balanced split possible given the distribution of the data. To train the regression models the raw scores were used as the training and testing targets, and the outputs were then quantized at the same thresholds as classification to compare accuracy results.

The distribution of situation awareness scores for Level 1 and Level 2 is shown in Figure 2. Level 1 scores have the widest spread, ranging from below 20 all the way to the 90s with the largest group of scores between 50 and 80. Level

2 scores, on the other hand, are skewed more towards higher scores; the lowest score is in the 40s, and the largest group of scores is between 70 and 90. Based on the difference in the distributions, the two levels should be quantized on different scores because each level has a different threshold of what is good. This can be explained intuitively with a driver in an automobile: a driver may be aware of the vehicles around them—which ones are changing lanes, speeding up behind them, etc.—but may not know their own exact speed or cardinal direction.

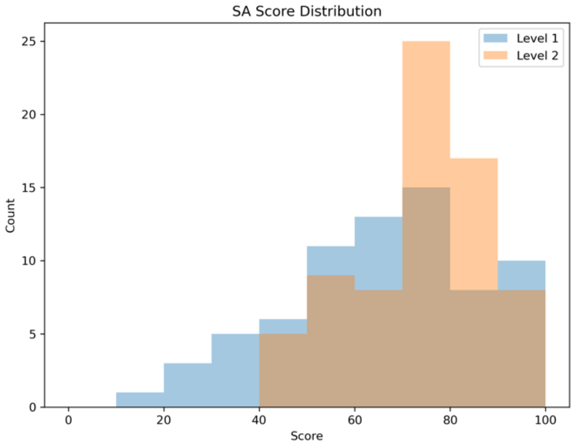

Figure 2. Distribution of situation awareness scores for Level 1 and Level 2.

SA Machine Learning Models

After feature extraction using the pretrained models, a new machine learning model is needed to transform the reduced dimensionality features into inferred SA scores. To achieve this, two types of models were trained for performance comparison: a neural network and a random forest. Both methods are popular and powerful analysis tools—but neural networks are characteristically trickier to optimize consistently. Therefore, random forest models are included alongside neural networks in our experiments for two main reasons: (1) we can determine if a neural network-based architecture is necessary to achieve good performance, or if a random forest-based architecture is sufficient; (2) we can use the feature importance values generated by the random forests to explore our hypotheses regarding the importance of gaze pattern accuracy and cognitive load for Level 1 and Level 2 situation awareness. The importance ranking gathered from the random forests are calculated based upon rankings. As such, the relative ranking of features can be used across most machine learning models. The raw importance values, however, are specific to the random forest model.

The neural network model has two layers: the first layer has an input of 208 dimensions with a 64-dimensional output, and the output layer has a single output dimension. The first layer has a ReLU activation, and the output layer has either a sigmoid activation for classification or a linear activation for regression. The model was trained for 50 epochs with an early stopping patience of 5 epochs using the Adam stochastic optimization method (Kingma and Ba, 2017) with a learning rate of 0.01 and default momentum cooling schedule values of β_1=0.9, β_2=0.999. The neural network was created and trained in Python using Tensorflow (Abadi, Agarwal, Barham, Brevdo, Chen, Citro, ... and Zheng, 2016) and Keras (Chollet, 2015). Hyper parameters for the neural network were found using randomized search via the Optuna package (Akiba, Sano, Yanase, Ohta, and Koyama, 2019). The random forest model (Breiman, 2001) was trained using both a classifier model and regressor model, as the neural network was. The random forest models for Level 1 used 100 trees and Gini impurity for splitting. The models for Level 2 used 1,000 trees, Gini impurity for splitting, and a max depth of 5 for each tree. The random forest models were fit using the Python package scikit-learn (Pedregosa, Varoquaux, Gramfort, Michel, Thirion, Grisel, ... and Duchesnay, 2011). Hyper parameters for the random forests were found using randomized search via the Optuna package (Akiba et al., 2019).

RESULTS

We employed two training tasks for evaluation: a regression trained model and classification model. Regression is ultimately a preferred task because it gives the designers of a training system more flexibility in defining what an acceptable level of SA should be for a given scenario. With a continuous score output, the distinction between pass and fail can be changed by the designers for different simulated scenarios, if desired. By comparing regression performance with classification performance, we can answer the following question: *Is it necessary to classify SA discretely, or does regression provide similar performance to classification?*

For all experiments, we used K-fold cross validation with 15 folds, with each fold stratified by the target quantized by the threshold corresponding to the given situation awareness level. That is, each fold contains approximately the same number of positive and negative examples. For Level 1 and Level 2, the thresholds were 65 and 77, respectively. All metrics shown are the performance on the held-out fold during cross validation. Level 1 and Level 2 have a 45/55% class split (as previously discussed), so accuracy and confusion matrices are meaningful criteria to use for evaluation. Table 2 aggregates the mean accuracy across all folds for all experiment configurations (a detailed per fold breakdown is described in a later section).

Table 2. Mean accuracy across all folds for each model and task.

Level 1 SA			**Level 2 SA**		
Task	**Model**	**Accuracy**	**Task**	**Model**	**Accuracy**
Regression	Random Forest	0.74	Regression	Random Forest	0.58
	Neural Network	0.75		Neural Network	0.66
Classification	Random Forest	0.77	Classification	Random Forest	0.63
	Neural Network	0.77		Neural Network	0.70

Based on the mean performance differences, we can compare the regression and classification performance presented. The difference in performance for Level 1 situation awareness is minimal, only a few percentage points worse for regression than classification. Additionally, both the random forest and neural network have comparable regression and classification performance for Level 1. However, neither model had a significant overall improvement over the other. For Level 1, then, regression produces good results that are comparable to classification for both random forests and neural networks.

This conclusion is not the case for Level 2. The random forest does a poor job at the regression task, achieving an accuracy of only 58% on quantized predictions. It does somewhat better at classification, achieving 63% accuracy. The neural network has significant improvement over the random forest for both regression and classification, achieving 66% accuracy for regression and 70% accuracy for classification. Based on the difference in performance for Level 2, the random forest does not seem to be able to create a model complex enough to perform well at either regression or classification, likely due to the more difficult task of predicting Level 2 situation awareness. The neural network is able to create a more complex model, so it can learn to more accurately predict Level 2.

Table 3. Per Model Confusions for Level 1 SA

Acc=0.74	**Rand Forest Regress., L1 SA**		Acc=0.77	**Rand Forest Class., L1 SA**	
Label: Fail	0.69	0.31	Label: Fail	0.68	0.32
Label: Pass	0.21	0.79	Label: Pass	0.19	0.81
	Model: Fail	Model: Pass		Model: Fail	Model: Pass
Acc=0.75	**Neural Net Regress., L1 SA**		Acc=0.77	**Neural Net Class., L1 SA**	
Label: Fail	0.79	0.21	Label: Fail	0.74	0.26
Label: Pass	0.28	0.72	Label: Pass	0.20	0.80
	Model: Fail	Model: Pass		Model: Fail	Model: Pass

Based on the confusion matrices for Level 1 SA (Table 3), the neural network model better predicted true negatives, while the random forest better predicted true positives. The random forest model has more false positives than false negatives, meaning the model tends to over predict scores. In contrast, the neural network has more false negatives than false positives, although this difference is less significant. Moreover, the neural network predicted more true negatives than the random forest by about 6 points. Additionally, the neural network predicted false negatives and false positives at approximately the same rate, whereas the random forest predicted significantly more false positives than false negatives. Compared to regression, the classification models performed only marginally better.

Based on the confusion matrices for Level 2 SA (Table 4), both the random forest and neural network models struggle to classify true negatives for Level 2 SA. Comparing the confusion matrices shows where the neural network has improved performance over the random forest. The neural network has a 15 point improvement in true negatives, with only 4 points worse in true positives. Because the neural network did not improve true positives, it had over twice as many false negatives than false positives compared to the random forest's small difference between false positives and false negatives. The neural network performs much better at classifying true positives than the random forest; interestingly, the neural network is able to predict true negatives well in the regression task, the opposite phenomenon. This is a general trend across all models: during classification, true positives are predicted more accurately than true negatives, and during regression true negatives are predicted more accurately than true positives. For Level 2, there is a significant drop in accuracy when using regression over classification. All models poorly predict true negatives, only correct predicting 60% of them. However, the neural network has a large improvement in predicting true positives, correctly predicting 81% of them. Therefore, the only acceptable performance achievable for Level 2 situation awareness is a neural network classifying the scores.

Table 4. Per Model Confusions for Level 2 SA

Acc=0.58	**Rand Forest Regress., L2 SA**		Acc=0.63	**Rand Forest Class., L2 SA**	
Label: Fail	0.63	0.37	Label: Fail	0.60	0.40
Label: Pass	0.44	0.56	Label: Pass	0.34	0.66
	Model: Fail	Model: Pass		Model: Fail	Model: Pass
Acc=0.66	**Neural Net Regress., L2 SA**		Acc=0.70	**Neural Net Class., L2 SA**	
Label: Fail	0.78	0.22	Label: Fail	0.60	0.40
Label: Pass	0.48	0.52	Label: Pass	0.19	0.81
	Model: Fail	Model: Pass		Model: Fail	Model: Pass

Detailed Per Fold Performance Results

To help elucidate the consistency of the performance across folds, we employ two metrics, mean squared error and accuracy, in swarm plots to show each result per fold. In each figure, the plots are swarm plots of the average accuracy and mean squared error for each fold, showing Level 1 and Level 2 on each plot.

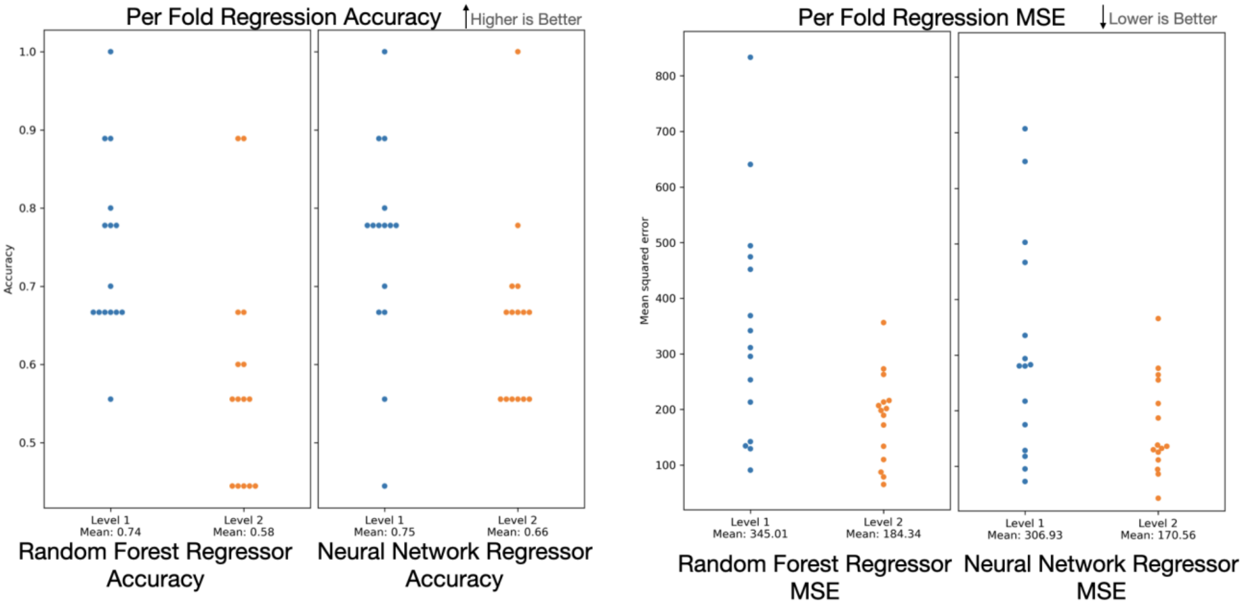

Figure 3. Per fold accuracy and mean squared error (MSE) for the regression models of level 1 and level 2 SA

Recall from Table 2 that, for Level 1 SA, the mean accuracy of the random forest performs on par with the neural network. In the swarm plot shown in Figure 3, however, one can see that the mean accuracy for the neural network is largely influenced by one fold's accuracy of around 45%. Two-thirds of the neural network folds have an accuracy of 70% or greater, while only around one-half of the random forest folds have an accuracy of 70% or greater. The neural network tends to have a larger variance in accuracy, but also has more high-scoring folds than the random forest. Examining the mean squared error, the neural network has an average score of around 40 points less than the random forest. The distribution of the scores for the random forest is somewhat uniform between 100 and 500, comprising 14 of the 15 folds. On the other hand, the neural network has tighter groups of scores around 100 and 300, with 10 folds less than 300 points. Thus, we can conclude that the neural network is preferrable to the random forest, even though the average performances are similar.

Recall from Table 2 that, for Level 2 SA, the difference in performance is much clearer, with preference to the neural network. The random forest has a mean accuracy of 58%, whereas the neural network has a mean accuracy of 66%, a difference of 8 points. The random forest had five folds with an accuracy of around 45%, whereas the lowest accuracy folds for the neural network was around 55% accuracy. The distributions of accuracy were almost identical, but the neural network had a distribution that was shifted about 10 points higher. The mean squared error distributions for both models were also quite similar, a mean difference of only around 14 points. The reason mean squared error is so much smaller in Level 2 than Level 1 while having a worse accuracy is due to the difference in score distributions, depicted in Figure 2. Level 2 SA scores are skewed much more towards 100, so the potential errors are smaller while being more difficult to predict.

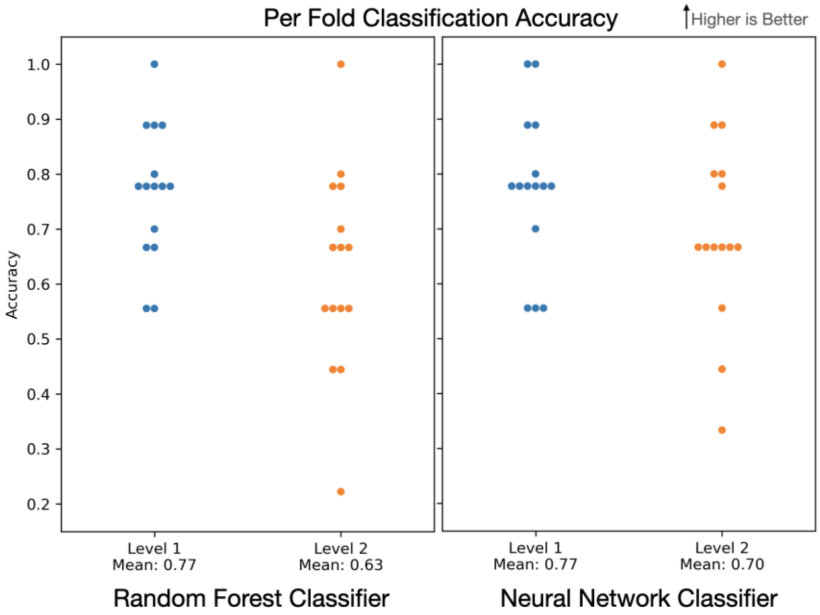

Figure 4. Per fold accuracy for the classification models.

Recall from Table 2 that, for Level 1 SA, the mean accuracy for the random forest is the same as the neural network. The per fold accuracies for these models are shown in Figure 4. Because the output of the classifier is "pass" and "fail" a mean squared error analysis is not possible. As seen in the swarm plots in Figure 4 the distribution of accuracy scores for Level 1 SA is similar, with the largest group of scores around 80%. The swarm plot results for Level 2 show a noticeable improvement in accuracy and confusions for the neural network compared to the random forest. The swarm plot in Figure 4 shows a large group of folds around 70% accuracy with the rest of the folds spread out, most of them at and above 80%. The distribution of scores for the random forest in Figure 4 is more uniformly distributed between 45% and 80%. This validates the conclusions from our initial analysis—the neural network classifier shows greater promise in classifying overall SA compared to the random forest.

Feature Importance

As discussed earlier, we previously hypothesized the importance of visual scan accuracy and cognitive load for Level 1 and Level 2 situation awareness. We hypothesized that visual scan accuracy would have the greatest importance in achieving Level 1 while cognitive load would have the greatest importance in achieving Level 2. To investigate these hypotheses, we used the feature importance generated by the fit random forests on the classification task for both Level 1 and Level 2. There were 208 feature importance rankings that mapped to the 208-dimensional input feature vectors. Using normalization of these rankings the sum of the importance equals 1. The first 96 features corresponded to gaze pattern, while the last 112 features corresponded to cognitive load. These two segments were separately analyzed, and the results marginally support our hypothesis. For Level 1 SA roughly 80% of the prediction is due to gaze features, while only 20% of the prediction is ascertained from cognitive load features. These importance values are not merely the result of an imbalance in the number of features for each modality; the gaze features make up 46% of the input vector yet provide 80% of the importance. Thus, there is a strong preference for gaze features in the Level 1 SA prediction. For Level 2 SA, the preference in features is more balanced. For Level 2 SA, 67% of the prediction is due to gaze features, and 33% is due to the cognitive load features. Thus, there is still a preference for gaze features, but not as strong as the Level 1 SA prediction models. This result validates our hypothesis regarding Level 1 situation awareness and provides evidence for our hypothesis for Level 2 SA, although gaze-based features clearly have an important role even for Level 2 SA prediction. Based on our training methodology, gaze pattern accuracy is the most important factor for both Level 1 and Level 2 SA.

CONCLUSION

In this research we evaluated random forests and neural networks in both regression and classification tasks for situation awareness. We find that we can accurately and reliably use biometric and eye tracking data to evaluate Level 1 SA in both regression and classification tasks, and we can accurately evaluate Level 2 SA using the same data in a classification task. To accomplish this, we utilize our previous research in gaze pattern classification and cognitive load assessment built on neural network models. These models create a strong feature extraction process through which the raw biometric and gaze data can be transformed into more meaningful features to train SA evaluation. Using these features, a neural network model produces the most consistently best-performing evaluation for both regression and classification. Level 1 SA can be evaluated to comparable mean results using both regression and classification (77% accuracy), while Level 2 SA requires a neural network classifier to produce good evaluation results (70% accuracy). For Level 1 SA, the gaze pattern features contribute more to the output than the cognitive load features, but this observation was seen to a lesser extent for Level 2 SA. This supports our initial hypotheses regarding feature importance that gaze has more influence on Level 1 SA queries. Based on these results, we conclude that accurately evaluating SA using data automatically collected from a pilot has great potential to improve the efficiency of training pilots. Future research should improve the accuracy of SA classification and verify that they are indeed diagnostic of performance. We are confident such milestone can be achieved, which will allow to apply real-time SA indices to optimize training flows.

ACKNOWLEDGMENTS

We would like to acknowledge the help of Major Justin Wilson, PhD, now a Professor at the United States Air Force Academy, for his assistance in creating standalone classifiers for cognitive load and scan pattern classification from previous work. This was instrumental in facilitating transfer learning for SA classification.

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., ... & Zheng, X. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019, July). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 2623-2631).

Baxter, J., Caruana, R., Mitchell, T., Pratt, L. Y., Silver, D. L., & Thurn, S. (1995). Post-NIPS* 95 Workshop on Transfer in Inductive Systems. Retrieved 2019-11-10, from http://socrates. acadiau. ca/courses/comp/dsilver/NIPS95_LTL/\\transfer. workshop. 1995. html.

Chollet, F. (2015). keras. github (2015).

Endsley, M. R. (1988, October). Design and evaluation for situation awareness enhancement. In *Proceedings of the Human Factors Society annual meeting* (Vol. 32, No. 2, pp. 97-101). Sage CA: Los Angeles, CA: Sage Publications.

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human factors*, *37*(1), 32-64.

Endsley, M. R. (2017). Direct measurement of situation awareness: Validity and use of SAGAT. In *Situational awareness* (pp. 129-156). Routledge.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology* (Vol. 52, pp. 139-183). North-Holland.

Hart, S. G. (2006, October). NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 50, No. 9, pp. 904-908). Sage CA: Los Angeles, CA: Sage Publications.

Khalili-Mahani, N., Assadi, A., Li, K., Mirgholami, M., Rivard, M. E., Benali, H., ... & De Schutter, B. (2020). Reflective and reflexive stress responses of older adults to three gaming experiences in relation to their cognitive abilities: mixed methods crossover study. *JMIR mental health*, *7*(3), e12388.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, *22*(10), 1345-1359.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830.

Roscoe, A. H., & Ellis, G. A. (1990). *A subjective rating scale for assessing pilot workload in flight: A decade of practical use*. ROYAL AEROSPACE ESTABLISHMENT FARNBOROUGH (UNITED KINGDOM).

Scielzo, S., Wilson, J. C., & Larson, E. C. (2020). Towards the development of an automated, real-time, objective measure of situation awareness for pilots. In *The Interservice/Industry Training, Simulation and Education Conference (I/ITSEC), Orlando, FL*.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Snyder, D., Garcia-Romero, D., Povey, D., & Khudanpur, S. (2017, August). Deep neural network embeddings for text-independent speaker verification. In *Interspeech* (pp. 999-1003).

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018, April). X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5329-5333). IEEE.

Wilson, J., Scielzo, S., Nair, S., & Larson, E. C. (2020). Automatic Gaze Classification for Aviators: Using Multi-task Convolutional Networks as a Proxy for Flight Instructor Observation. *International Journal of Aviation, Aeronautics, and Aerospace*, *7*(3), 7.

Wilson, J. C., Nair, S., Scielzo, S., & Larson, E. C. (2021). Objective Measures of Cognitive Load Using Deep Multi-Modal Learning: A Use-Case in Aviation. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *5*(1), 1-35.